Spring 2014

Nebraska State Accountability (NeSA)

Reading, Mathematics, and Science

Alternate Assessment

Technical Report

August 2014

Prepared by Data Recognition Corporation

DATA RECOGNITION

DRC

CORPORATION

2014 NEBRASKA STATE ACCOUNTABILITY (NeSA) ALTERNATE ASSESSMENT TECHNICAL REPORT

TABLE OF CONTENTS

1. BACKGROUND

2. ITEM AND TEST DEVELOPMENT

1. Background

1.1 Purpose and Organization of This Report

This report documents the technical aspects of the 2014 Nebraska Alternate Assessment Reading (NeSA-AAR), Mathematics (NeSA-AAM), and Nebraska Science (NeSA-AAS) operational tests, along with the NeSA-AAR, NeSA-AAM and NeSA-AAS embedded field tests, covering details of item and test development process, administration procedures, and psychometric methods and summaries.

1.2 Background of the Nebraska State Accountability (NeSA)

Previous Nebraska Alternate Assessments: Prior to 2009, Alternate Assessments were not required. Districts had the ability to locally administer Alternate Assessments to students of their districts.

Purpose of the NeSA: Legislative Bill 1157 passed by the 2008 Nebraska Legislature (http://www.legislature.ne.gov/laws/statutes.php?statute=79-760.03) required a single statewide assessment of the Nebraska academic content standards for reading, mathematics, science, and writing in Nebraska's K-12 public schools. The new assessment system was named NeSA (Nebraska State Accountability), with NeSA-AAR for alternate reading assessments, NeSA-AAM for alternate mathematics, NeSA-AAS for alternate science. The alternate assessments in reading and mathematics were administered in grades 3-8 and 11; science was administered in grades 5, 8, and 11.

The NeSA-Alternate Assessment (NeSA-Alt) consists entirely of multiple choice items and are administered in a paper pencil format. In January 2009, the NDE contracted with Data Recognition Corporation (DRC) to support the Department of Education with the administration, record keeping, and reporting of statewide student assessment and accountability.

Phase-In Schedule for NeSA Alternate Assessment: The NDE prescribed the regular and the Alternate assessments starting in the 2009-2010 school year to be phased in as shown in Table 1.1. The state intends to use the expertise and experience of in-state educators to participate, to the maximum extent possible, in the design and development of the new statewide assessment system.

Table 1.1: NeSA Regular and Alternate Assessment Administration Schedule

Subject	Administration Year		Grades
	Field Test	Operational	
Reading	2009	2010	3 through 8 plus high school
Mathematics	2010	2011	3 through 8 plus high school
Science	2011	2012	5, 8 and 11

Advisory Committees: Legislative Bill 1157 added a governor-appointed Technical Advisory Committee (TAC) with three nationally recognized experts in educational assessment, one Nebraska administrator, and one Nebraska teacher. The TAC reviewed the development plan for the NeSA

Alternate Assessment, and provided technical advice, guidance, and research to help the NDE make informed decisions regarding standards, assessment, and accountability.

1.3 Administration

The NeSA-Alt assessments are administered to students individually. The test administrator reads a prepared script for each item. As part of the assessment, the administrator may read the items multiple times and each student responds in their primary mode of communication. Test administrators record each response on the answer sheet. Students are able to utilize a full range of allowable accommodations that are detailed in documentation from the Nebraska Department of Education. If it becomes clear that a student is unable to respond to questions, the test administrator is required to record this on the answer sheet. Students who were administrated the test but unable to respond count as participants but receive a zero score.

2. Item and Test Development

2.1 Content Standards

In April of 2008, the Nebraska Legislature passed into state law Legislative Bill 1157. This action changed previous provisions related to standards, assessment, and reporting. Specific to standards, the legislation stated:

- The State Board of Education shall adopt measurable academic content standards for at least the grade levels required for statewide assessment. The standards shall cover the content areas of reading, writing, mathematics, and science. The standards adopted shall be sufficiently clear and measurable to be used for testing student performance with respect to mastery of the content described in the state standards.
- The State Board of Education shall develop a plan to review and update standards for each content area every five years.
- The State Board of Education shall review and update the standards in reading by July 1, 2009, the standards in mathematics by July 1, 2010, and these standards in all other content areas by July 1, 2013.

The Nebraska Language Arts Standards are the foundation for NeSA-AAR. This assessment instrument is comprised of items that address standards for grades 3–8 and 12. The standards are assessed at grade-level with the exception of grade 12. The grade 12 standards are assessed on the NeSA-AAR tests at grade 11. The reading standards for each grade are represented in items that are distributed between two reporting categories: Vocabulary and Comprehension. The Vocabulary standards include word structure, context clues, and semantic relationships. The Comprehension standards include author's purpose, elements of narrative text, literary devices, main idea, relevant details, text features, genre, and generating questions while reading.

The mathematics component of the NeSA-AAM is composed of items that address indicators in grades 3–8 and high school. The standards are assessed at grade level with the exception of high school. The high school standards are assessed on the NeSA-AAM at grade 11. The assessable standards for each grade level are distributed among the four reporting categories: Number Sense Concepts, Geometric/Measurement Concepts, Algebraic Concepts, and Data Analysis/Probability Concepts.

The science component of the NeSA-AAS is composed of items that address indicators in grade-band strands 3–5, 6–8, and 9–12. The NeSA-AAS assesses the standards for each grade-band strand at a specific grade: 3-5 strand at grade 5, 6–8 strand at grade 8, and 9–12 strand at grade 11. The assessable standards for each grade level are distributed among the four reporting categories: Inquiry, The Nature of Science, and Technology; Physical Science; Life Science; and Earth and Space Sciences.

The NeSA-Alt are based on the same set of content standards that were extended by a team of special education specialists. The extended indicators detail underlying skills that students need to master prior to attaining mastery of the full standard. The NeSA-Alt are aligned to the extended indicators.

2.2 Test Blueprints (Table of Specifications)

The test blueprints, or Table of Specifications (TOS), for each assessment include lists of all the standards, organized by reporting categories. The test blueprints also contain the Depth of Knowledge (DOK) level ranges assigned to each standard and the range of test items to be part of the assessment by extended indicator. The NeSA-AAR test blueprint (Appendix A) was originally developed and approved in fall 2009. The NeSA-AAM test blueprint (Appendix B) was originally developed and approved in fall 2010. The NeSA-AAS test blueprint (Appendix C) was originally developed and approved in fall 2011.

As part of the maturation of the NeSA-Alt program, NDE undertook to clarify the TOS in fall 2013 based on a careful examination of the overall pool of items within the NeSA-Alt item bank and the characteristics of the previous successful operational administrations. As a result, clarifications were made to all three TOS to better reflect the historical content of the NeSA-Alt program, and the clarified TOS were posted to NDE's website in advance of the 2013-2014 school year. It is important to point out that the clarifications made to the TOS bring the NeSA-Alt TOS into alignment with the actual historical NeSA-Alt test blueprints but did not change the breadth or depth of the content assessed within the actual NeSA-Alt program.

2.3 Multiple-Choice Items

Each assessment incorporates multiple-choice (MC) items to assess the content standards. Students are required to select a correct answer from three response choices with a single correct answer. Each MC item is scored as right or wrong and has a value of one raw score point. MC items are used to assess a variety of skill levels in relation to the tested standards.

2.4 Item Development and Review

The most significant considerations in the item and test development process are: aligning the items to the grade level extended indicators; determining the grade-level appropriateness; DOK; estimated difficulty level; and determining style, accuracy, and correct terminology. In addition, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and *Universal Design* (Thompson, Johnstone, & Thurlow, 2002) guided the following steps in the item development process:

- Analyze the grade-level extended indicators and test blueprints.
- Analyze item specifications and style guides.
- Select qualified item writers.
- Develop item-writing workshop training materials.
- Train Nebraska educators to write items.

- Write items that match the standards, are free of bias, and address fairness and sensitivity concerns.
- Conduct and monitor internal item reviews and quality processes.
- Select and assemble items for field testing.
- Field test items, score the items, and analyze the data.
- Review items and associated statistics after field testing, including bias statistics.
- Update item bank.

Item Writer Training: The test items were written by Nebraska educators who were recommended for the process by an administrator. Three criteria were considered in selecting the item writers: educational role, geographic location, and experience with item writing.

Prior to developing items for NeSA-Alt, a cadre of item writers was trained with regard to:

- Nebraska content standards and test blueprints;
- cognitive levels, including Depth of Knowledge (DOK);
- principles of Universal Design;
- skill-specific and balanced test items for the grade level;
- developmentally appropriate structure and content;
- item-writing technical quality issues;
- bias, fairness, and sensitivity issues; and
- style considerations and item specifications.

Item Writing: To ensure that all test items met the requirements of the approved target content test blueprint and were adequately distributed across subcategories and levels of difficulty, item writers were asked to document the following specific information as each item was written:

- Alignment to the Nebraska Standards: There must be a high degree of match between a particular question and the standard it is intended to measure. Item writers were asked to clearly indicate which extended indicator each item was measuring.

- Appropriate Grade Level, Item Context, and Assumed Student Knowledge: Item writers were asked to consider the conceptual and cognitive level of each item. They were asked to review each item to determine whether or not the item was measuring something that was important and could be successfully taught and learned in the classroom.

- MC Item Options and Distractor Rationale: Writers were instructed to make sure that each item had only one clearly correct answer. Item writers submitted the answer key with the item. All distractors were plausible choices that represented common errors and misconceptions in student reasoning.

- Face Validity and Distribution of Items Based upon DOK: Writers were asked to classify the DOK of each item, using a model based on Norman Webb's work on four DOK categories:

recall, skill/concept, strategic thinking, and extended thinking (Webb, 2002). The NeSA-Alt items are classified based on DOK stages, subsets of the four categories. The stages include: responding, reproducing, recalling and basic reasoning.

- Readability: Writers were instructed to pay careful attention to the readability of each item to ensure that the focus was on the concepts; not on reading comprehension of the item. Resources writers used to verify the vocabulary level were the *EDL Core Vocabularies* (Taylor, Frackenpohl, White, Nieroroda, Browning, & Brisner, 1989) and the *Children's Writer's Word Book* (Mogilner, 1992). In addition, every test item was reviewed by grade-level experts. They reviewed each item from the perspective of the students they teach, and they determined the validity of the vocabulary used.
- Grammar and Structure for Item Stems and Item Options: All items were written to meet technical quality, including correct grammar, syntax, and usage in all items, as well as parallel construction and structure of text associated with each MC item.

Item Review: Throughout the item development process, independent panels of reading content experts and special education specialists reviewed the items. The following guidelines for reviewing assessment items were used during each review process.

A quality item should:

- have only one clear correct answer and contain answer choices that are reasonably parallel in length and structure;
- have a correctly assigned content code (item map);
- measure one main idea or problem;
- measure the objective or curriculum content standard it is designed to measure;
- be at the appropriate level of difficulty;
- be simple, direct, and free of ambiguity;
- make use of vocabulary and sentence structure that is appropriate to the grade level of the student being tested;
- be based on content that is accurate and current;
- when appropriate, contain stimulus material that are clear and concise and provide all information that is needed;
- when appropriate, contain graphics that are clearly labeled;
- contain answer choices that are plausible and reasonable in terms of the requirements of the question, as well as the students' level of knowledge;
- contain distractors that relate to the question and can be supported by a rationale;
- reflect current teaching and learning practices in the content area; and
- be free of gender, ethnic, cultural, socioeconomic, and regional stereotyping bias.

Following each review process, the item writer group and the item review panel discussed suggestions for revisions related to each item. Items were revised only when both groups agreed on the proposed change.

Editorial Review of Items: After items were written and reviewed, the NDE test development specialists reviewed each item for item quality, making sure that the test items were in compliance with guidelines for clarity, style, accuracy, and appropriateness for Nebraska students. Additionally, DRC test development content experts worked collaboratively with the NDE to review and revise the items prior to field testing to ensure highest level of quality possible.

Universally Designed Assessments: Universally designed assessments allow participation of the widest possible range of students and result in valid inferences about performance of all students who participate and are based on the premise that each child in school is a part of the population to be tested, and that testing results should not be affected by disability, gender, race, or English language ability (Thompson, Johnstone, & Thurlow, 2002). The NDE and DRC are committed to the development of items and tests that are fair and valid for all students. At every stage of the item and test development process, procedures ensure that items and tests are designed and developed using the elements of universally designed assessments that were developed by the National Center on Educational Outcomes (NCEO).

Federal legislation addresses the need for universally designed assessments. The *No Child Left Behind Act* (Elementary and Secondary Education Act) requires that each state must "provide for the participation in [statewide] assessments of all students" [Section 1111(b)(3)(C)(ix)(l)]. Both Title 1 and IDEA regulations call for universally designed assessments that are accessible and valid for all students including students with disabilities and students with limited English proficiency. The NDE and DRC recognize that the benefits of universally designed assessments not only apply to these groups of students, but to all individuals with wide-ranging characteristics.

The NDE test development team and Nebraska item writers have been trained in the elements of Universal Design as it relates to developing large-scale statewide assessments. Additionally, the NDE and DRC partner to ensure that all items meet the Universal Design requirements during the item review process.

After a review of research relevant to the assessment development process and the principles of Universal Design (Center for Universal Design, 1997), NCEO has produced seven elements of Universal Design as they apply to assessments (Thompson, Johnstone, & Thurlow, 2002).

Inclusive Assessment Population

> When tests are first conceptualized, they need to be thought of in the context of who will be tested. If the test is designed for state, district, or school accountability purposes, the target population must include every student who will participate in accountability through an alternate assessment. The NDE and DRC are fully aware of increased demands that statewide assessment systems must include and be accountable for ALL alternate students.

Precisely Defined Constructs

An important function of well-designed assessments is that they actually measure what they are intended to measure. The NDE item writers and DRC carefully examine what is to be tested and design items that offer the greatest opportunity for success within those constructs. Just as universally designed architecture removes physical, sensory, and cognitive barriers to all types of people in public and private structures, universally designed assessments must remove all non-construct-oriented cognitive, sensory, emotional, and physical barriers.

Accessible, Non-biased Items

The NDE conducts both internal and external review of items and test specifications to ensure that they do not create barriers because of lack of sensitivity to disability, cultural, or other subgroups. Items and test specifications are developed by a team of individuals who understand the varied characteristics of items that might create difficulties for any group of students. Accessibility is incorporated as a primary dimension of test specifications, so that accessibility is woven into the fabric of the test rather than being added after the fact.

Amenable to Accommodations

Even though items on universally designed assessments will be accessible for most students, there will still be some students who continue to need accommodations for the alternate test. Thus, another essential element of any universally designed assessment is that it is compatible with accommodations and a variety of widely used adaptive equipment and assistive technology. NDE and DRC work to ensure that state guidelines on the use of accommodations are compatible with the assessment being developed.

Simple, Clear, and Intuitive Instructions and Procedures

Assessment instructions should be easy to understand, regardless of a student’s experience, knowledge, language skills, or current cognitive level. Directions and questions need to be in simple, clear, and understandable language. Knowledge questions that are posed within complex language certainly invalidate the test if students cannot understand how they are expected to respond to a question.

Maximum Readability and Comprehensibility

A variety of guidelines exist to ensure that text is maximally readable and comprehensible. These features go beyond what is measured by readability formulas. Readability and comprehensibility are affected by many characteristics, including student background, sentence difficulty, organization of text, and others. All of these features are considered as the NDE develops the text of assessments.

Plain language is a concept now being highlighted in research on assessments. Plain language has been defined as language that is straightforward and concise. The following strategies for editing text to produce plain language are used during the NDE’s editing process:

- Reduce excessive length.
- Use common words.
- Avoid ambiguous words.
- Avoid irregularly spelled words.
- Avoid proper names.
- Avoid inconsistent naming and graphic conventions.
- Avoid unclear signals about how to direct attention.
- Mark all questions.
- Maximum legibility.

Legibility is the physical appearance of text, the way that the shapes of letters and numbers enable people to read text easily. Bias results when tests contain physical features that interfere with a student's focus on or understanding of the constructs that test items are intended to assess. DRC works closely with the NDE to develop a style guide that includes dimensions of style that are consistent with universal design.

DOK: Interpreting and assigning DOK levels to both objectives within standards and assessment items is an essential requirement of alignment analysis. Four levels of DOK are used for this analysis. The NeSA-Alt assessments include items written at levels 1 and 2. Levels 3 and 4 items are not included due to the test being comprised of only MC items and the cognitive level of students taking the alternate assessments. In addition, the NeSA-Alt items are classified based on DOK stages—subsets of the four DOK levels. The stages include responding, reproducing, recalling at DOK 1, and basic reasoning at DOK 2.

Reading Level 1-Stage 1: Responding to Discourse Materials

Level 1-Stage 1 requires students to display the ability to respond to or indicate, or acknowledge text or discourse related features. Some examples that represent, but do not constitute all of, Level1-Stage 1 performance are:

- Student demonstrates the ability to attend to pictures/symbols/objects pertinent to a story
- Students display attention to people, surroundings, or materials.
- Student attends while teacher reads.

Reading Level 1-Stage 2: Reproduce Discourse Related Materials

Level 1-Stage 2 requires students to display the ability to copy, replicate, repeat, re-enact, mirror, or match text or discourse related features. Some examples that represent, but do not constitute all of, Level 1-Stage 2 performance are:

- Students match pictures and/or words that depict emotions such happy, sad, or angry.
- Students match printed words to objects.

Reading Level 1-Stage 3: Recalls Information about Discourse Related Materials

Level 1-Stage 3 requires the ability to recite or recall facts or information. Involves the ability to distinguish between text-based or discourse features. Some examples that represent, but do not constitute all of, Level 1-Stage3 performance are:

- Students demonstrate understanding or new words or passages by making connections with personal experience via speech, writing, signs, or assistive device.
- Students retell information taken from printed materials.
- Students answer who, what and where questions about a story.

Reading Level 2-Stage 4: Basic Reasoning

Level 2-Stage 4 requires processing beyond recall and observation. This requires both comprehension and subsequent processing of text. It also involves ordering, classifying text as well as identifying patterns, relationships, and main points. Some examples that represent, but do not constitute all of, Level 2-Stage 4 performance are:

- Students correct grammar mistakes in a reading selection.
- Students summarize the main idea of paragraph.
- Students identify the author's purpose for writing a brief passage.

Mathematics Level 1-Stage 1: Responding to Mathematical Materials

Level 1-Stage 1 requires the ability to respond to, indicate, or acknowledge mathematical features. Some examples that represent, but do not constitute all of, Level1-Stage 1 performance are:

- Students are able to recognize that there is a difference in patterns.
- Students respond to math ideas using appropriate vocabulary.

Mathematics Level 1-Stage 2: Reproduce Mathematical Features

Level 1-Stage 2 requires the ability to copy, replicate, repeat, re-enact, mirror, or match mathematical features. Some examples that represent, but do not constitute all of, Level 1-Stage 2 performance are:

- Students will write numbers accurately in a variety of contexts.
- Student accurately sort basic shapes into groups
- Student is able to accurately identify location terms when prompted (i.e., next to, between, over, under).

Mathematics Level 1-Stage 3: Recalls Information about Mathematical Features

Level 1-Stage 3 requires students to recall or observe facts, definitions, terms. It also involves simple one-step procedures. The stage also includes computing simple algorithms (e.g., sum, quotient). Some examples that represent, but do not constitute all of, Level 1-Stage3 performance are:

- Students locate a pattern in order to solve a problem
- Students measures using feet and yards.
- Students use a calculator or concrete objects to add and subtract.

Mathematics Level 2-Stage 4: Basic Reasoning

Level 2-Stage 4 requires students to make decisions of how to approach a problem. This may require students to compare, classify, organize, estimate or order data. This also typically involves two-step procedures. Some examples that represent, but do not constitute all of, Level 2-Stage 4 performance are:

- Student reads problem and determines operation to solve the problem.
- Student selects geometric figure from group of figures based on the definition of the geometric figure.
- Student determines how to solve for unknown value in equation or inequality and then selects solution.

Science Level 1-Stage 1: Responding to Scientific Features

Level 1-Stage 1 requires the ability to respond to or indicate or acknowledge scientific features. Some examples that represent, but do not constitute all of, Level1-Stage 1 performance are:

- Students attend to a teacher conducting scientific inquiry.
- Students respond to science ideas using appropriate vocabulary.

Science Level 1-Stage 2: Reproduce Scientific Features

Level 1-Stage 2 requires the ability to copy, replicate, repeat, re-enact, mirror, or match scientific ideas. Some examples that represent, but do not constitute all of, Level 1-Stage 2 performance are:

- Students copy figure of animal with distinguishing features.
- Student matches numbers on measuring devices.
- Student is able to accurately match descriptions of living and nonliving objects to visual representations.

Science Level 1-Stage 3: Recalls Information about Scientific Features

Level 1-Stage 3 requires students to recall or observe facts, definitions, terms. It also involves simple one-step procedures. The stage also requires a demonstration of a rote response, use of a well-known formula, or follow a set procedure (like a recipe), or preform a clearly defined series of steps. Some examples that represent, but do not constitute all of, Level 1-Stage3 performance are:

- Students recall or recognize a fact, term, or property.
- Students identify the correct measuring device to perform a task.
- Students perform a routine safety procedure.

Science Level 2-Stage 4: Basic Reasoning

Level 2-Stage 4 requires students to make decisions of how to approach a question or problem. This may require students to classify, organize, estimate, make observations or collect and order data. This also typically involves two-step procedures. Some examples that represent, but do not constitute all of, Level 2-Stage 4 performance are:

- Students make observations and collect data.
- Students organize and display data in tables, graphs, and charts.
- Students describe and explain examples and non-examples of science concepts.

2.5 Item Banking

Prior to 2013, NDE exclusively maintained an item bank that provided a repository of item image, history, statistics, and usage. The item bank included a record of all newly created items together with item data from each item field test. It also included all data from the operational administration of the items. Within the item bank, NDE:

- updated the information after each administration;
- updated the information with newly developed items;
- monitored the content to ensure an appropriate balance of items aligned with content standards, goals, and objectives;
- monitored item history statistics; and
- monitored the content for an appropriate balance of DOK levels.

In preparation for the 2014 NeSA Alternate test administration, NDE transitioned the item bank to DRC. DRC now maintains the alternate item bank in their system known as IDEAS, and it now functions as a repository of item image, history, statistics, and usage for the NeSA-Alt. IDEAS includes a record of all newly created items together with item data from each item field test. It also includes all data from the operational administration of the items. Within IDEAS, DRC:

- updates the Nebraska item bank after each administration;
- updates the Nebraska item bank with newly developed items;
- monitors the Nebraska item bank to ensure an appropriate balance of items aligned with content standards, goals, and objectives;
- monitors item history statistics; and
- monitors the Nebraska item bank for an appropriate balance of DOK levels.

2.6 The Operational Form Construction Process

The Spring 2014 operational forms were constructed in Lincoln, Nebraska in late August and early September of 2013. The forms were constructed by a team of specialists representing special education, the Nebraska Department of Education, and DRC testing experts. Training was provided collaboratively by NDE and DRC for the forms construction process.

Since this was the first opportunity for DRC to participate in the NeSA-Alt operational form construction process with NDE, NDE provided DRC with background information related to the NeSA-Alt program in advance of the meetings in Lincoln, explaining the philosophy of the program and its goals. Prior to arrival in Lincoln, DRC Test Development content specialists reviewed the test blueprints and the item pool to ensure that there was alignment between the items and the indicators, including the number of items per standard for each content-area test.

The specialists were provided with an overview of the psychometric guidelines and targets for operational forms construction. The foremost guideline was for item content to match the test blueprint (Table of Specifications) for the given content. The point-biserial correlation guideline was to be greater than 0.35 (with a requirement for no point-biserial correlation less than zero). In addition, the average target *p*-value for each test was to be about 0.65. The overall summary of the actual approved *p*-value and biserial of the forms is provided in the summary table later in this document. Below is the psychometric guidelines followed for item selection.

Psychometric Guidelines for Item Selection for a New Assessment[1]

The main headings are more or less in order of precedence. This effectively means that content and reliability (*IIa and IIb*) define the pool of eligible items, from which items are selected based on p-value to match a target. *Guideline* is used here in the sense of *guiding principle*, not in the sense of *strict rule*. It is often, perhaps typically, necessary to deviate from these principles for a few items. There is no guideline for what *a few items* means.

I. ***Item content: match the blue print.***

II. Item-Total Correlation: (aka, item reliability or, for MC, point-biserial correlation)

   a. Absolutely no correlations less than zero. This is a requirement, not a guideline.
   b. **Ideally, for multiple-choice (MC), correlation should be greater than *0.35*.**
      i. A low correlation indicates there is a *smart* way to get the item wrong or a *not-smart* way to get it right.
      ii. The lower value, the weaker the item.

III. P-value for Correct Response on MC

   a. Target **mean percent correct about 65%** plus or minus a couple percent.
   b. Ideally, all items **greater than 40% and less than 85%**.
   c. For an existing assessment, the target mean percent correct should approximate the past forms.

[1] For an existing assessment, the target mean percent correct should approximate the past forms.

DRC Test Development specialists printed a copy of each item card, with accompanying item characteristics, image, and psychometric data. Test Development specialists verified the accuracy of each item card, making sure that the item image has its correct item characteristics. Test Development specialists carefully reviewed each item card's psychometric data to ensure it is complete and reasonable. The item cards were compiled in binders and sorted by standard and indicator.

The NDE and DRC also checked to see that each item met technical quality for well-crafted items, including:

- only one correct answer,
- wording that is clear and concise,
- grammatical correctness,
- appropriate item complexity and cognitive demand,
    - appropriate range of difficulty,
    - appropriate depth-of-knowledge alignment,
- aligned with principles of Universal Design, and
- free of any content that might be offensive, inappropriate, or biased (content bias).

NDE representatives and DRC Test Development specialists made initial grade-level selections of the items, known as the "pull list," to be included on the 2014 operational forms. The goal was for the first pull of the items to meet the Table of Specification (TOS) guidelines and psychometric guidelines specific to each content area. As items were selected, the unique item codes were entered using software into a form building template (PerForm) which contained the item pool with statistics and item characteristics. The template automatically calculated the *p*-value, biserial, number of items per indicator and standard, number of items per DOK level, and distribution of answer key as items were selected for each grade. As items were selected, the item characteristics (key, DOK, and alignment to indicator) were verified.

Review of the Items and Test Forms: At every stage of the test development process, the match of the item to the content standard was reviewed and verified, since establishing content validity is one of the most important aspects in the legal defensibility of a test. As a result, it is essential that an item selected for a form link directly to the content curriculum standard and performance standard to which it is measuring. NDE specialists verified all items against their classification codes and item maps, both to evaluate the correctness of the classification and to ensure that the given task measures what it purports to measure.

2.7 Reading Assessment

Test Design: The NeSA-AAR operational test includes operational items and field test items. The form pools contained 25 operational items and 16 field test items.

Table 2.7.1 Reading 2014 Operational Test

Grade	Total No. of MC Core Items	No. of Embedded FT Items per Form	Total Items per Form	Total No. of Equivalent FT Forms	Total Core Points	Total No. of MC Items Added to the Bank
3	25	8	33	2	25	16
4	25	8	33	2	25	16
5	25	8	33	2	25	16
6	25	8	33	2	25	16
7	25	8	33	2	25	16
8	25	8	33	2	25	16
11	25	8	33	2	25	16

Equating Design: Spring 2014 was the fifth operational administration of the NeSA-AAR. Approximately 20–40% of the assessment was constructed from items field tested from Spring 2009–2013. The approximate remaining 60–80% of the assessment was constructed from an overlap of items from the 2010–2013 operational (core) item positions from the Spring 2010–2013 operational forms.

In addition to the operational items, each student received 8 selected field test items. Equating was accomplished by anchoring on the operational items and calibrating the field test items concurrently.

2.8 Mathematics Assessment

Test Design: The NeSA-AAM operational test includes operational items and field test items. The form pools contained 25 or 30 operational items (depending on the grade) with 16 field test items.

Table 2.8.1 Mathematics 2014 Operational Test

Grade	Total No. of MC Core Items	No. of Embedded FT Items per Form	Total Items per Form	Total No. of Equivalent FT Forms	Total Core Points	Total No. of MC Items Added to the Bank
3	25	8	33	2	25	16
4	30	8	38	2	30	16
5	30	8	38	2	30	16
6	30	8	38	2	30	16
7	30	8	38	2	30	16
8	30	8	38	2	30	16
11	30	8	38	2	30	16

Equating Design: Spring 2014 was the fourth operational administration of the NeSA-AAM. Approximately 20–40% of the assessment was constructed from items field tested from Spring 2010–

2013. The approximate remaining 60–80% of the assessment was constructed from an overlap of items from the 2011–2013 operational (core) item positions from the 2011–2013 operational forms.

In addition to the operational items, each student received 8 selected field test items. Equating was accomplished by anchoring on the operational items and calibrating the field test items concurrently.

2.9 Science Assessment

Test Design: The NeSA-AAS operational test includes operational and field test items. Depending on grade, the form pools contained 25 or 30 operational items (depending on the grade) with 16 field test items.

Table 2.9.1 Science 2014 Operational Test

Grade	Total No. of MC Core Items	No. of Embedded FT Items per Form	Total Items per Form	Total No. of Equivalent FT Forms	Total Core Points	Total No. of MC Items Added to the Bank
5	25	8	33	2	25	16
8	25	8	33	2	25	16
11	30	8	38	2	30	16

Equating Design: Spring 2014 was the third operational administration of the NeSA-AAS. Approximately 20–60% of the assessment was constructed from items field tested in Spring 2011–2013. The approximate remaining 40–80% of the assessment was constructed from an overlap of items from the 2012–2013 operational (core) item positions from the 2012–2013 operational forms.

In addition to the operational items, each student received 8 field test items. Equating was accomplished by anchoring on the operational items and calibrating the field test items concurrently.

3. STUDENT DEMOGRAPHICS AND ACCOMMODATIONS

Gender, ethnicity, food program status (FRL), Limited English Proficiency/English Language Learners (LEP/ELL) status, and accommodation status data was collected for all students who participated and attempted the 2014 NeSA-Alt. This summary of student demographics by grade and content area is provided in Tables 3.1.1– 3.1.7. These tables show that for each grade, around 300 students took the assessment. Of those students across grades, over half (at least 60%) are males, over half are white, and less than one fifth are Hispanic. Among the students across grades, over half are eligible for FRL, and almost all are non-LEP/ELL. In terms of the test accommodations, there are over half of the students across grade and content area that report at least one type of accommodation (see row 'Total' for 'Accommodation' in the table). Across all grades, the 'Timing/Schedule/Setting' is the most utilized accommodation, followed by the 'Response' and 'Content Presentation'.

Table 3.1.1 Grade 3 NeSA-Alt Summary Data: Demographics and Accommodations

Grade 3		Reading		Mathematics	
		Count	%	Count	%
All Students		282	100.0	276	100.0
Gender	Female	100	35.5	99	35.9
	Male	182	64.5	177	64.1
Race/Ethnicity	American Indian/Alaska Native	11	3.9	11	4.0
	Asian	6	2.1	6	2.2
	Black	20	7.1	20	7.2
	Hispanic	64	22.7	60	21.7
	Native Hawaiian or other Pacific Islander	0	0.0	0	0.0
	White	168	59.6	166	60.1
	Two or More Races	13	4.6	13	4.7
Food Program	Yes	175	62.1	172	62.3
	No	98	34.8	95	34.4
LEP/ELL	Yes	3	1.1	3	1.1
	No	279	98.9	273	98.9
Accommo-dations	Content Presentation	138	48.9	132	47.8
	Response	149	52.8	141	51.1
	Timing/Schedule/Setting	196	69.5	186	67.4

Grade 3		Reading		Mathematics	
		Count	%	Count	%
	Direct Linguistic Support with Test Directions	4	1.4	4	1.4
	Direct Linguistic Support with Content and Test items	4	1.4	3	1.1
	Indirect Linguistic Support	2	0.7	2	0.7
	Total	197	69.9	186	67.4

Table 3.1.2 Grade 4 NeSA-Alt Summary Data: Demographics and Accommodations

Grade 4		Reading		Mathematics	
		Count	%	Count	%
All Students		332	100.0	333	100.0
Gender	Female	109	32.8	111	33.3
	Male	223	67.2	222	66.7
Race/Ethnicity	American Indian/Alaska Native	10	3.0	10	3.0
	Asian	5	1.5	5	1.5
	Black	23	6.9	23	6.9
	Hispanic	72	21.7	72	21.6
	Native Hawaiian or other Pacific Islander	0	0.0	0	0.0
	White	211	63.6	212	63.7
	Two or More Races	11	3.3	11	3.3
Food Program	Yes	188	56.6	186	55.9
	No	137	41.3	139	41.7
LEP/ELL	Yes	2	0.6	3	0.9
	No	330	99.4	330	99.1
Accommo-dations	Content Presentation	165	49.7	161	48.3
	Response	169	50.9	164	49.2
	Timing/Schedule/Setting	217	65.4	214	64.3
	Direct Linguistic Support with Test Directions	0	0.0	2	0.6

Grade 4		Reading		Mathematics	
		Count	%	Count	%
	Direct Linguistic Support with Content and Test items	0	0.0	1	0.3
	Indirect Linguistic Support	1	0.3	2	0.6
	Total	220	66.3	216	64.9

Table 3.1.3 Grade 5 NeSA-Alt Summary Data: Demographics and Accommodations

Grade 5		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
All Students		326	100.0	331	100.0	325	100.0
Gender	Female	106	32.5	112	33.8	109	33.5
	Male	220	67.5	219	66.2	216	66.5
Race/Ethnicity	American Indian/Alaska Native	4	1.2	4	1.2	4	1.2
	Asian	3	0.9	3	0.9	3	0.9
	Black	38	11.7	38	11.5	37	11.4
	Hispanic	52	16.0	53	16.0	54	16.6
	Native Hawaiian or other Pacific Islander	0	0.0	0	0.0	0	0.0
	White	218	66.9	222	67.1	216	66.5
	Two or More Races	11	3.4	11	3.3	11	3.4
Food Program	Yes	182	55.8	184	55.6	181	55.7
	No	132	40.5	135	40.8	132	40.6
LEP/ELL	Yes	1	0.3	1	0.3	1	0.3
	No	325	99.7	330	99.7	324	99.7
Accommo-dations	Content Presentation	164	50.3	167	50.5	162	49.8
	Response	160	49.1	165	49.8	160	49.2
	Timing/Schedule/Setting	196	60.1	198	59.8	191	58.8
	Direct Linguistic Support with Test Directions	2	0.6	1	0.3	4	1.2
	Direct Linguistic Support with Content and Test items	1	0.3	1	0.3	2	0.6

Grade 5		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
	Indirect Linguistic Support	2	0.6	1	0.3	3	0.9
	Total	201	61.7	206	62.2	198	60.9

Table 3.1.4 Grade 6 NeSA-Alt Summary Data: Demographics and Accommodations

Grade 6		Reading		Mathematics	
		Count	%	Count	%
All Students		319	100.0	317	100.0
Gender	Female	123	38.6	127	40.1
	Male	196	61.4	190	59.9
Race/Ethnicity	American Indian/Alaska Native	5	1.6	5	1.6
	Asian	3	0.9	3	0.9
	Black	37	11.6	36	11.4
	Hispanic	67	21.0	63	19.9
	Native Hawaiian or other Pacific Islander	0	0.0	0	0.0
	White	194	60.8	196	61.8
	Two or More Races	13	4.1	14	4.4
Food Program	Yes	186	58.3	181	57.1
	No	128	40.1	131	41.3
LEP/ELL	Yes	4	1.3	4	1.3
	No	315	98.7	313	98.7
Accommo-dations	Content Presentation	166	52.0	169	53.3
	Response	159	49.8	160	50.5
	Timing/Schedule/Setting	200	62.7	200	63.1
	Direct Linguistic Support with Test Directions	1	0.3	1	0.3
	Direct Linguistic Support with Content and Test items	0	0.0	0	0.0
	Indirect Linguistic Support	0	0.0	0	0.0

Grade 6		Reading		Mathematics	
		Count	%	Count	%
	Total	205	64.3	205	64.7

Table 3.1.5 Grade 7 NeSA-Alt Summary Data: Demographics and Accommodations

Grade 7		Reading		Mathematics	
		Count	%	Count	%
All Students		334	100.0	332	100.0
Gender	Female	114	34.1	116	34.9
	Male	220	65.9	216	65.1
Race/Ethnicity	American Indian/Alaska Native	7	2.1	7	2.1
	Asian	6	1.8	6	1.8
	Black	40	12.0	39	11.7
	Hispanic	53	15.9	56	16.9
	Native Hawaiian or other Pacific Islander	0	0.0	0	0.0
	White	211	63.2	208	62.7
	Two or More Races	17	5.1	16	4.8
Food Program	Yes	187	56.0	186	56.0
	No	139	41.6	138	41.6
LEP/ELL	Yes	2	0.6	2	0.6
	No	332	99.4	330	99.4
Accommo-dations	Content Presentation	141	42.2	141	42.5
	Response	147	44.0	145	43.7
	Timing/Schedule/Setting	180	53.9	176	53.0
	Direct Linguistic Support with Test Directions	1	0.3	1	0.3
	Direct Linguistic Support with Content and Test items	1	0.3	1	0.3
	Indirect Linguistic Support	1	0.3	1	0.3
	Total	183	54.8	181	54.5

Table 3.1.6 Grade 8 NeSA-Alt Summary Data: Demographics and Accommodations

Grade 8		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
All Students		337	100.0	344	100.0	333	100.0
Gender	Female	133	39.5	135	39.2	130	39.0
	Male	204	60.5	209	60.8	203	61.0
Race/Ethnicity	American Indian/Alaska Native	2	0.6	2	0.6	2	0.6
	Asian	14	4.2	14	4.1	14	4.2
	Black	31	9.2	33	9.6	31	9.3
	Hispanic	48	14.2	48	14.0	48	14.4
	Native Hawaiian or other Pacific Islander	0	0.0	0	0.0	0	0.0
	White	225	66.8	230	66.9	221	66.4
	Two or More Races	17	5.0	17	4.9	17	5.1
Food Program	Yes	183	54.3	186	54.1	178	53.5
	No	146	43.3	150	43.6	147	44.1
LEP/ELL	Yes	0	0.0	0	0.0	0	0.0
	No	337	100	344	100.0	333	100.0
Accommo-dations	Content Presentation	151	44.8	156	45.3	149	44.7
	Response	151	44.8	156	45.3	146	43.8
	Timing/Schedule/Setting	185	54.9	192	55.8	180	54.1
	Direct Linguistic Support with Test Directions	0	0.0	0	0.0	0	0.0
	Direct Linguistic Support with Content and Test items	1	0.3	1	0.3	1	0.3
	Indirect Linguistic Support	0	0.0	0	0.0	0	0.0
	Total	191	56.7	197	57.3	186	55.9

Table 3.1.7 Grade 11 NeSA-Alt Summary Data: Demographics and Accommodations

Grade 11		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
All Students		297	100.0	304	100.0	296	100.0
Gender	Female	123	41.4	125	41.1	123	41.6
	Male	174	58.6	179	58.9	173	58.4
Race/Ethnicity	American Indian/Alaska Native	7	2.4	7	2.3	7	2.4
	Asian	4	1.3	4	1.3	4	1.4
	Black	37	12.5	37	12.2	37	12.5
	Hispanic	34	11.4	34	11.2	34	11.5
	Native Hawaiian or other Pacific Islander	0	0.0	0	0.0	0	0.0
	White	202	68.0	207	68.1	201	67.9
	Two or More Races	13	4.4	15	4.9	13	4.4
Food Program	Yes	148	49.8	154	50.7	148	50.0
	No	148	49.8	149	49.0	147	49.7
LEP/ELL	Yes	2	0.7	2	0.7	2	0.7
	No	295	99.3	302	99.3	294	99.3
Accommo-dations	Content Presentation	95	32.0	100	32.9	96	32.4
	Response	107	36.0	109	35.9	104	35.1
	Timing/Schedule/Setting	162	54.5	165	54.3	161	54.4
	Direct Linguistic Support with Test Directions	3	1.0	2	0.7	3	1.0
	Direct Linguistic Support with Content and Test items	3	1.0	3	1.0	3	1.0
	Indirect Linguistic Support	1	0.3	1	0.3	1	0.3
	Total	169	56.9	174	57.2	167	56.4

4. CLASSICAL ITEM STATISTICS

This chapter provides an overview of the most familiar item-level statistics obtained from classical (traditional) item analysis: item difficulty, item discrimination, distractor distribution, and omits or blanks. The following results pertain only to operational NeSA-Alt items (i.e., those items that contributed to a student's total test score). Rasch item statistics are discussed in Chapter Five, and test-level statistics are found in Chapter Six. The statistics provide information about the quality of the items based on student responses in an operational setting. The following sections provide descriptions of the item summary statistics found in Appendices F, G, and H.

4.1 ITEM DIFFICULTY

Item difficulty (*p*-value) is the proportion of examinees in the sample who answered the item correctly. For example, if an item has a *p*-value of 0.89, it means 89 percent of the students answered the item correctly. Relatively lower values correspond to more difficult items and those that have relatively higher values correspond to easier items. Items that are either very hard or very easy provide little information about student differences in achievement. On a standards-referenced test like the NeSA-Alt, a test development goal is to include a wide range of item difficulties. Typically, test developers target *p*-values in the range of 0.30 to 0.90. Mathematically, information is maximized and standard errors minimized when the *p*-value equals 0.50. Experience suggests that multiple choice items are effective when the student is more likely to succeed than fail and it is important to include a range of difficulties matching the distribution of student abilities (Wright & Stone, 1979). Occasionally, items that fall outside the desired range can be justified for inclusion when the educational importance of the item content or the desire to measure students with very high or low achievement override the statistical considerations. Summary *p*-value information across all grades for each content area is shown in Tables 4.1.1 – 4.1.3. In general, most of the items fall into the *p*-value range of 0.4 to 0.9, which is appropriate for a criterion-referenced assessment.

Table 4.1.1 Summary of Traditional Item Proportion Correct for NeSA-AAR Operational Items

	Item Proportion Correct											
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	Mean	Total
3	0	0	0	0	3	5	7	10	0	0	0.653	25
4	0	0	0	0	4	5	11	5	0	0	0.610	25
5	0	0	0	0	1	5	15	3	1	0	0.641	25
6	0	0	0	0	0	5	12	6	2	0	0.666	25
7	0	0	0	0	2	4	10	7	2	0	0.668	25
8	0	0	0	1	1	4	12	5	2	0	0.641	25
11	0	0	0	0	0	3	10	8	4	0	0.702	25

Table 4.1.2 Summary of Traditional Item Proportion Correct for NeSA-AAM Operational Items

	Item Proportion Correct											
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	Mean	Total
3	0	0	0	0	3	6	9	4	3	0	0.648	25
4	0	0	0	1	2	8	10	9	0	0	0.625	30
5	0	0	0	0	3	7	14	6	0	0	0.634	30
6	0	0	0	0	3	4	13	7	3	0	0.660	30
7	0	0	0	0	1	10	8	8	3	0	0.654	30
8	0	0	0	1	3	10	8	6	2	0	0.623	30
11	0	0	0	0	4	8	7	8	3	0	0.643	30

Table 4.1.3 Summary of Traditional Item Proportion Correct for NeSA-AAS Operational Items

	Item Proportion Correct											
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	Mean	Total
5	0	0	0	0	3	4	14	4	0	0	0.632	25
8	0	0	0	0	0	8	12	5	0	0	0.636	25
11	0	0	0	0	1	6	10	11	2	0	0.678	30

4.2 ITEM-TOTAL CORRELATION

Item-total correlation describes the relationship between performance on the specific item and performance on the entire form. For the NeSA-Alt tests, Pearson's product-moment correlation coefficient between item scores and test scores is used to indicate this relationship. For MC items, the statistic is typically referred to as *point-biserial correlation*. This index indicates an item's ability to differentiate between high and low achievers (i.e., item discrimination power). It is expected that students with high ability (i.e., those who perform well on the NeSA-Alt overall) would be more likely to answer any given NeSA-Alt item correctly, while students with low ability (i.e., those who perform poorly on the NeSA-Alt overall) would be more likely to answer the same item incorrectly. However, an interaction can exist between item discrimination and item difficulty. Items answered correctly (or incorrectly) by a large proportion of examinees (i.e., the items have extreme *p*-values) can have reduced power to discriminate and thus can have lower correlations.

The correlation coefficient can range from -1.0 to +1.0. If the aforementioned expectation is met (high-scoring students tend to get the item right while low-scoring students do not), the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., well above zero), meaning the item is a good discriminator between high- and low-ability students. Items with negative correlations are flagged and referred to Test Development as possible mis-keys. Mis-keyed items will be corrected and rescored prior to computing the final item statistics. Negative correlations can also indicate problems with the item content, structure, or students' opportunity to learn. Items with point-biserial values of less than 0.2 are flagged and referred to content specialists for review before being considered for use on future forms. As seen below in Tables 4.2.1 – 4.2.3, no items in the 2014 NeSA-Alt tests have negative point-biserial correlations and most are above 0.20, indicating good item discrimination.

Table 4.2.1 Summary of Point-biserial Correlations for NeSA-AAR

	Item Point-biserial Correlation							
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	Total
3	0	0	0	2	3	10	10	25
4	0	0	0	2	8	8	7	25
5	0	0	1	1	3	12	8	25
6	0	0	1	1	9	7	7	25
7	0	0	0	1	5	3	16	25
8	0	0	0	0	5	7	13	25
11	0	0	0	1	3	5	16	25

Table 4.2.2 Summary of Point-biserial Correlations for NeSA-AAM

	Item Point-biserial Correlation							
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	Total
3	0	0	1	1	5	8	10	25
4	0	0	1	0	2	14	13	30
5	0	0	0	1	5	11	13	30
6	0	0	2	4	10	9	5	30
7	0	0	0	3	4	11	12	30
8	0	0	0	4	5	9	12	30
11	0	0	0	3	2	17	8	30

Table 4.2.3 Summary of Point-biserial Correlations for NeSA-AAS

	Item Point-biserial Correlation							
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	Total
5	0	0	0	2	2	8	13	25
8	0	0	1	1	4	7	12	25
11	0	0	1	2	2	6	19	30

4.3 PERCENT SELECTING EACH RESPONSE OPTION

This index indicates the effectiveness of each distractor. In general, one expects the correct response to be the most attractive, although this need not hold for unusually challenging items. This statistic for the correct response option is identical to the *p*-value when considering MC items with a single correct response. Please see the detailed summary statistics for each grade and content area in Appendices F, G, and H.

4.4 POINT-BISERIAL CORRELATIONS OF RESPONSE OPTIONS

This index describes the relationship between selecting a response option for a specific item and performance on the entire test. The correlation between an incorrect answer and total test performance should be negative. The desired pattern is strong positive values for the correct option and strong negative values for the incorrect options. Any other pattern indicates a problem with the item or with the key. These patterns would imply a high ability way to answer incorrectly or a low ability way to answer correctly. Examples of these situations could be an item with an ambiguous or misleading distractor that was attractive to high-performing examinees or an item that depended on experience outside of instruction that was unrelated to ability. This statistic for the correct option is identical to the item-total correlation for MC items. Please see the detailed summary statistics for each grade and content area in Appendices F, G, and H.

4.5 PERCENT OF STUDENTS OMITTING AN ITEM

This statistic is useful for identifying problems with testing time and test layout. If the omit percentage is large for a single item, it could indicate a problem with the layout or content of an item. For example, students tend to skip items with wordy stems or that otherwise appear difficult or time consuming. While there is no hard and fast rule for what *large* means, and it varies with groups and ages of students, five percent omits is often used as a preliminary screening value.

Detailed results of the item analyses for the NeSA-AAR operational items are presented in Appendix F. Detailed results of the item analyses for the NeSA-AAM operational items are presented in Appendix G. Detailed results of the item analyses for the NeSA-AAS operational items are presented in Appendix H. Based on these analyses, items were selected for review if the *p*-value was less than 0.25 and the *item-total correlation* was less than 0.2. Items were identified as probable mis-keys if the *p*-value for the correct response was less than one of the incorrect responses and the *item-total correlation* was negative.

5. RASCH ITEM CALIBRATION

The particular item response theory (IRT) model used for the NeSA-Alt is based on the work of Georg Rasch. Rasch models have had a long-standing presence in applied testing programs and have been the methodology used to calibrate NeSA-Alt items in recent history. IRT has several advantages over classical test theory, so it has become the standard procedure for analyzing item response data in large-scale assessments. However, IRT models make a number of strong assumptions related to dimensionality, local independence, and model-data fit. Resulting inferences derived from any application of IRT rests strongly on the degree to which the underlying assumptions are met.

Generally, item calibration is the process of assigning a difficulty-parameter estimate to each item on an assessment so that all items are placed onto a common scale. This chapter briefly introduces the Rasch model, reports the results from evaluations of the adequacy of the Rasch assumptions, and summarizes Rasch item statistics for the 2014 NeSA-AAR, NeSA-AAM, and NeSA-AAS assessment.

5.1 DESCRIPTION OF THE RASCH MODEL

The Rasch rating scale model was used to calibrate the NeSA-Alt items. All NeSA-Alt assessment contains only MC items. According to the Rasch model, the probability of answering an item correctly is based on the difference between the ability of the student and the difficulty of the item. The Rasch model places both student ability and item difficulty (estimated in terms of log-odds, or logits) on the same continuum. When the model assumptions are met, the Rasch model provides estimates of a person's ability that are independent of the items employed in the assessment and conversely, estimates item difficulty independently of the sample of examinees (Rasch, 1960; Wright & Panchapakesan, 1969). (As noted in Chapter Four, interpretation of item *p*-values confounds item difficulty and student ability.) Appendix I provides a more detailed overview of Rasch Measurement.

5.2 CHECKING RASCH ASSUMPTIONS

Since the Rasch model was the basis of all calibration, scoring, and scaling analyses associated with the NeSA-Alt, the validity of the inferences from these results depends on the degree to which the assumptions of the model were met and how well the model fits the test data. Therefore, it is important to check these assumptions. This section evaluates the dimensionality of the data, local item independence, and item fit. It should be noted that only operational items were analyzed since they are the basis of student scores.

Unidimensionality: Rasch models assume that one dominant dimension determines the difference among students' performances. Principal Components Analysis (PCA) can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify whether any other dominant component(s) exist among the items. If any other dimensions are found, the unidimensionality assumption would be violated.

Tables 5.2.1, 5.2.2, and 5.2.3 present the PCA results for the reading, mathematics, and science assessments, respectively. The results include the eigenvalues and the percentage of variance explained for up to five components with eigenvalues greater than one. As can been seen in Table 5.2.1, the primary dimension for NeSA-AAR explained about 25 percent to 30 percent of the total variance across Grades 3–8 and 11. The eigenvalues of the second dimension ranged from 1.4 to 2.9. This indicates that the second dimension accounted for only 1.4 to 2.9 units out of about 37 units of total variance. Similar patterns are observed for the Mathematics and the Science test. Overall, the PCA suggests that there is one clearly dominant dimension for each NeSA-Alt assessment.

Table 5.2.1 NeSA-AAR Results from PCA

Grade	Component	Eigenvalue	Explained Variance
3	1	10.7	29.9%
	2	2.0	5.6%
	3	1.8	4.9%
	4	1.5	4.2%
	5	1.4	4.0%
4	1	11.6	28.0%
	2	2.2	5.4%
	3	1.6	3.9%
	4	1.6	3.8%
	5	1.6	3.8%
5	1	11.9	28.3%
	2	2.5	6.0%
	3	2.2	5.2%
	4	1.5	3.5%
	5	1.4	3.4%
6	1	9.7	24.5%
	2	2.9	7.4%
	3	1.8	4.5%
	4	1.6	4.0%
	5	1.3	3.4%
7	1	10.0	25.0%
	2	2.7	6.7%
	3	1.8	4.6%
	4	1.6	3.9%
	5	1.6	3.9%
8	1	11.2	27.1%
	2	2.3	5.7%
	3	1.7	4.2%
	4	1.6	3.8%
	5	1.4	3.5%
11	1	12.0	28.6%
	2	2.2	5.1%
	3	1.8	4.2%
	4	1.7	4.0%
	5	1.5	3.7%

Table 5.2.2 NeSA-AAM Results from PCA

Grade	Component	Eigenvalue	Explained Variance
3	1	10.2	29.0%
	2	2.0	5.6%
	3	1.7	4.8%
	4	1.6	4.5%
	5	1.5	4.2%
4	1	8.0	24.2%
	2	2.2	6.8%
	3	1.6	4.9%
	4	1.6	4.8%
	5	1.4	4.4%
5	1	8.1	24.5%
	2	2.4	7.1%
	3	1.6	4.8%
	4	1.5	4.5%
	5	1.4	4.1%
6	1	8.6	25.5%
	2	2.0	5.9%
	3	1.6	4.7%
	4	1.5	4.4%
	5	1.4	4.2%
7	1	11.5	31.5%
	2	2.4	6.5%
	3	1.6	4.3%
	4	1.5	4.0%
	5	1.4	3.8%
8	1	11.3	31.2%
	2	1.9	5.2%
	3	1.6	4.5%
	4	1.5	4.1%
	5	1.4	3.8%
11	1	11.6	31.6%
	2	2.0	5.3%
	3	1.6	4.4%
	4	1.4	3.8%
	5	1.4	3.7%

Table 5.2.3 NeSA-AAS Results from PCA

Grade	Component	Eigenvalue	Explained Variance
5	1	10.1	28.8%
	2	2.0	5.8%
	3	1.7	5.0%
	4	1.6	4.4%
	5	1.4	3.9%
8*	1	8.2	24.8%
	2	2.4	7.3%
	3	1.4	4.3%
	4	1.4	4.1%
	5		
11	1	13.0	30.2%
	2	2.2	5.2%
	3	1.8	4.2%
	4	1.7	3.9%
	5	1.5	3.6%

*Only four components with eigenvalues greater than one were extracted.

Local Independence: Local independence (LI) is a fundamental assumption of IRT. No relationship should exist between examinees' responses to different items after accounting for the abilities measured by a test. Many indicators of LI are framed by the form of local independence proposed by McDonald (1979) that the conditional covariances of all pairs of item responses, conditioned on the abilities, are required to be equal to zero.

Residual item correlations provided in WINSTEPS for each item pair were used to assess local dependence among the NeSA-Alt items. Three types of residual correlations are available in WINSTEPS: raw, standardized, and logit. It should be noted that the raw score residual correlation essentially corresponds to Yen's $Q3$ index, a popular LI statistic. The expected value for the $Q3$ statistic is approximately $-1/(k-1)$ when no local dependence exists, where k is test length (Yen, 1993). Thus, the expected $Q3$ values should be approximately -0.04 for the NeSA-Alt tests (since most of the NeSA-Alt tests had more than 25 core items). Index values that are greater than 0.20 indicate a degree of local dependence that probably should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS was used for these analyses. Tables 5.2.4 – 5.2.6 show the summary statistics—mean, SD, minimum, maximum, and several percentiles (P10, P25, P50, P75, P90)—for all the residual correlations for each test. The total number of item pairs (N) and the number of pairs with the residual correlations greater than 0.20 are also reported in this table. The mean residual correlations were slightly negative and the values were close to -0.04. The vast majority of the correlations were very small, suggesting local item independence generally holds for the NeSA-Alt reading, mathematics, and science assessments.

Table 5.2.4 Summary of Item Residual Correlations for NeSA-AAR

Statistics	3	4	5	6	7	8	11
N	300	300	300	300	300	300	300
Mean	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
SD	0.08	0.08	0.08	0.07	0.08	0.07	0.07
Minimum	-0.23	-0.26	-0.21	-0.23	-0.27	-0.21	-0.21
P10	-0.13	-0.13	-0.14	-0.12	-0.13	-0.11	-0.11
P25	-0.09	-0.1	-0.1	-0.09	-0.09	-0.08	-0.08
P50	-0.04	-0.04	-0.04	-0.04	-0.05	-0.04	-0.04
P75	0.01	0.01	0.01	0.01	0.01	0	0
P90	0.06	0.06	0.07	0.05	0.07	0.05	0.05
Maximum	0.24	0.27	0.18	0.18	0.27	0.17	0.15
>0.20	1	1	0	0	2	0	0

Table 5.2.5 Summary of Item Residual Correlations for NeSA-AAM

	Mathematics						
Statistics	3	4	5	6	7	8	11
N	300	435	435	435	435	435	435
Mean	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
SD	0.07	0.07	0.08	0.09	0.09	0.07	0.07
Minimum	-0.24	-0.22	-0.24	-0.24	-0.26	-0.25	-0.26
P10	-0.13	-0.12	-0.13	-0.14	-0.13	-0.12	-0.12
P25	-0.09	-0.08	-0.1	-0.1	-0.09	-0.08	-0.08
P50	-0.04	-0.03	-0.04	-0.04	-0.04	-0.04	-0.04
P75	0.01	0.02	0.02	0.02	0.02	0.02	0.01
P90	0.06	0.06	0.08	0.09	0.08	0.06	0.06
Maximum	0.2	0.22	0.24	0.33	0.27	0.23	0.32
>0.20	1	1	1	4	4	1	1

Table 5.2.6 Summary of Item Residual Correlations for NeSA-AAS

	Science		
Statistics	5	8	11
N	300	300	435
Mean	-0.04	-0.04	-0.03
SD	0.08	0.08	0.08
Minimum	-0.24	-0.21	-0.24
P10	-0.13	-0.14	-0.13
P25	-0.1	-0.09	-0.08
P50	-0.04	-0.05	-0.03
P75	0.02	0.01	0.03
P90	0.06	0.07	0.07
Maximum	0.16	0.19	0.21
>0.20	0	0	2

Item Fit: WINSTEPS provides two item fit statistics (infit and outfit) for evaluating the degree to which the Rasch model predicts the observed item responses. Each fit statistic can be expressed as a mean square (MnSq) statistic or on a standardized metric (Zstd with mean = 0 and variance = 1).

MnSq values are more oriented toward practical significance, while Zstd values are more oriented toward statistical significance. Though both are informative, the Zstd values are very likely too sensitive to the large sample sizes. In this situation it is recommended that the Zstd values be ignored if the MnSq values are acceptable (Linacre, 2009).

The outfit statistic tends to be affected more by unexpected responses far from the person, item, or rating scale category measure (i.e., it is more sensitive to outlying, off-target, and low information responses). The infit statistic tends to be affected more by unexpected responses close to the person, item, or rating scale category measure (i.e., informative, on-target responses). Some researchers contend that extreme infit values are a greater threat to the measurement process than extreme outfit since most tests intend to measure the on-target population rather than extreme outliers.

The expected MnSq value is 1.0 and can range from 0 to infinity. Deviation in excess of the expected value can be interpreted as noise or lack of fit between the items and the model. Values lower than the expected value can be interpreted as item redundancy or overfitting items (too predictable and/or too much redundancy), and values greater than the expected value indicate underfitting items (too unpredictable and/or too much noise). Rules of thumb regarding “practically significant” MnSq values vary. More conservative users might prefer items with MnSq values that range from 0.8 to 1.2. Others believe reasonable test results can be achieved with values from 0.5 to 1.5. In the results below, values outside of 0.7 to 1.3 are given practical importance.

Table 5.2.7 presents the summary statistics of infit and outfit mean square statistics for the NeSA-Alt reading, mathematics, and science tests, including the mean, SD, and minimum and maximum values. The number of items within the range of [0.7, 1.3] is also reported in Table 5.2.7. As can be seen, the mean values for both fit statistics were close to 1.00 for all tests. Most of the items had infit values falling in the range of [0.7, 1.3]. Though more outfit values fell outside this range than infit values, it is not surprising given that the test is designed to measure the on-target population than extreme outliers. Overall, these results indicate that the Rasch model fits the NeSA-Alt item data well.

Table 5.2.7 Summary of Infit and Outfit Mean Square Statistics for 2014 NeSA-Alt Tests

		Infit Mean Square					Outfit Mean Square				
		Mean	SD	MIN	MAX	[0.7, 1.3]	Mean	SD	MIN	MAX	[0.7, 1.3]
Reading	3	0.97	0.18	0.62	1.35	21/25	0.93	0.30	0.39	1.50	16/25
	4	1.01	0.18	0.72	1.37	24/25	0.99	0.30	0.51	1.67	15/25
	5	1.01	0.19	0.71	1.48	23/25	1.02	0.37	0.51	2.27	17/25
	6	1.00	0.18	0.72	1.46	24/25	1.01	0.43	0.46	2.70	17/25
	7	0.99	0.25	0.64	1.55	19/25	0.99	0.42	0.37	1.94	11/25
	8	1.01	0.18	0.72	1.31	24/25	0.99	0.30	0.56	1.51	15/25
	11	0.98	0.23	0.67	1.57	22/25	0.97	0.40	0.44	2.17	15/25
Mathematics	3	1.01	0.16	0.70	1.41	23/25	0.97	0.32	0.34	1.59	16/25
	4	1.02	0.20	0.70	1.63	28/30	0.99	0.32	0.47	1.93	22/30
	5	1.02	0.16	0.76	1.35	28/30	0.99	0.25	0.61	1.54	22/30
	6	1.02	0.16	0.77	1.36	28/30	0.98	0.31	0.56	1.75	19/30
	7	1.00	0.19	0.73	1.48	28/30	0.97	0.34	0.45	1.84	20/30
	8	1.00	0.18	0.66	1.36	27/30	0.96	0.31	0.44	1.56	15/30
	11	1.00	0.14	0.62	1.35	28/30	0.98	0.26	0.47	1.83	25/30
Science	5	1.03	0.22	0.76	1.47	21/25	1.04	0.39	0.60	1.83	16/25
	8	1.03	0.21	0.76	1.54	22/25	1.04	0.39	0.58	1.88	16/25
	11	0.97	0.24	0.53	1.53	24/30	0.91	0.40	0.30	2.07	12/30

5.3 RASCH ITEM STATISTICS

Item calibration was implemented via WINSTEPS 3.80.1 program (Linacre, 2014). The characteristics of calibration samples are reported in Chapter Three. These samples only include the students who attempted the tests. All omits (no response) and multiple responses (more than one response selected) were scored as incorrect answers (coded as 0s) for calibration.

As noted earlier, the Rasch model expresses item difficulty (and student ability) in units referred to as *logits* rather than on the proportion-correct metric. Large negative logits represent easier items while large positive logits represent more difficult items. Logits have an interval scale, meaning that two items with logits of 0.0 and +1.0 (respectively) are the same distance apart (in difficulty) as two items with logits of +3.0 and +4.0.

Appendices J, K, L, and M report the Rasch calibration summaries and logit difficulties for all the operational items. Table 5.3.1 summarizes the Rasch logit difficulties of the operational items on each test. The minimum and maximum values and standard deviations suggest that the NeSA-Alt items covered a relatively wide range of difficulties. It is important to note that the logit difficulty values presented have not been linked to a common scale of measurement. Therefore, the relative magnitude of the statistics across subject areas and grades cannot be compared. The item pool was then updated with the item statistics.

Table 5.3.1 Summary of Rasch Item Difficulties for NeSA-AAR, NeSA-AAM, and NeSA-AAS

	Grade	N	Mean	SD	Min	Max
Reading	**3**	25	0.05	0.63	-0.95	1.14
	4	25	0.25	0.61	-0.84	1.34
	5	25	0.11	0.51	-1.00	0.98
	6	25	0.31	0.45	-0.56	1.29
	7	25	-0.06	0.83	-1.74	1.40
	8	25	0.52	0.80	-1.27	1.90
	11	25	-0.22	0.64	-1.43	1.40
Mathematics	**3**	25	-0.27	0.92	-2.21	1.29
	4	30	0.07	0.70	-1.23	1.18
	5	30	-0.1	0.77	-1.46	1.60
	6	30	0.04	0.71	-1.56	1.62
	7	30	-0.16	0.71	-1.52	1.03
	8	30	0.01	0.75	-1.47	1.39
	11	30	-0.36	0.80	-1.88	0.85
Science	**5**	25	-0.99	0.67	-2.29	0.07
	8	25	-1.08	0.56	-2.54	0.23
	11	30	-1.21	0.67	-2.73	0.37

6. EQUATING AND SCALING

As discussed earlier in Chapter 2, the 2014 test forms were constructed with items that were either field tested, or used operationally, on a previously administered NeSA-Alt test. The NeSA-Alt assessment is constructed each year allowing each NeSA-Alt assessment to be different from the previous year's assessment. To ensure that all forms for a given grade and content area provide comparable scores, and to ensure the passing standards across different administrations are equivalent, the new operational items need to be placed on the bank scale via equating to bring the 2014 NeSA-Alt raw-score-to-Rasch-ability scale to the previous operational scale. When the new 2014 NeSA-Alt tests are placed on the bank's scale, the resulting scale scores for the new test form will be the same as the scale scores of the previous operational form such that students performing at the same level of (underlying) achievement should receive the same score (i.e., scale score). The resulting scale scores will be used for score reporting and performance level classification. Once operational items are equated, field test items are then placed on the bank scale and are then ready for future operational use.

This chapter begins with a summary of the entire NeSA-Alt equating procedures. This is followed by a scaling analysis that transforms raw scores to scale scores that represent the same skill level on every test form. Some summary results of the state scale score performance are also provided.

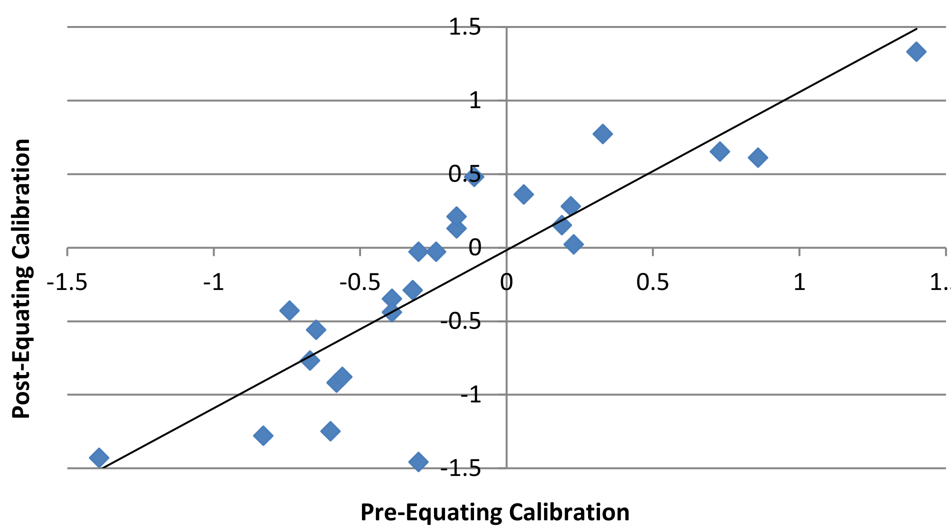
6.1 EQUATING

The equating design employed for NeSA-Alt is often referred to as a common-item nonequivalent groups (CINEG) design, which uses a set of common anchor items to adjust for differences in test difficulty across years. If the item properties (i.e., difficulty) calibrated from the previous administrations hold true for the current student population, the whole set of the 2014 NeSA-Alt operational items can serve as the linking set such that conversions from raw to scale scores can be established prior to the time when the new test is administered operationally. This is often referred to as the pre-equating process because it establishes the raw-to-scale-score conversion based on the previously calibrated item difficulties (and can happen prior to the administration of the operational test).The most appealing feature of the pre-equating process, when applicable, is its ability to facilitate immediate score reporting for tests which have tight reporting windows.

Unlike the 2013 NeSA-Alt tests, the entire set of items in the 2014 NeSA-Alt tests has been placed to the bank scale, which makes the pre-equating solution possible for the 2014 NeSA-Alt tests. However, it may not be wise to assume that the operational items maintain their relative difficulty across years because the same item can have different properties in different years because of changes in the item's position or changes in the students' experiences. Therefore, once the 2014 operational test data is available, DRC Psychometric Services staffers together with NDE evaluated the item difficulty equivalence via the so-called 'post-equating check' procedure to identify items that have significant difficulty changes from the bank scale. If no unstable item is identified, the 2014 equating process would be suggested to follow a pre-equating solution. On the other hand, if an item or items are found to be unstable, the difficulties of these items need to be calibrated using the 2014 operational test data. The set of 2014 operational items with those identified items excluded will be used as the set to link the 2014 test to the bank scale. This equating process is known as the post-equating process because the equating happens after the administration of the operation test and the raw-to-scale-score conversion is generated based on the operational test data.

As the post-equating check procedures, DRC Psychometric Services staffers evaluated the item difficulty equivalence by comparing the old banked item calibration (called pre-calibration) with a new unanchored calibration of the 2014 data (called post-calibration). The evaluations were conducted for each grade and content area, using both visual graphing and statistical methods. The post-calibrated item difficulties (logits) were plotted against the pre-calibration for each grade and content area (see Appendices N – P). Ideally, these scatter plots should have a strong linear trend. Items straying from the trend line did not perform in the same way in both years. Below is an example of pre- and post-calibration plots for the 2014 NeSA-AAR test (Grade 11). Graphically, there is one apparent outlier item that strays far from the trend line. It is located at the lower middle range that, for some reason, became much easier for the population of this year. All the other items fall more or less on the linear trend line, indicating consistent performance in both years.

Figure 6.1.1: NeSA-AAR Grade 11 Pre- and Post-Calibrations



DRC Psychometric Services staff examined the robust *Z* statistic (Huynh, 2000; Huynh & Rawls, 2009), the correlations between the pre- and post-calibrated item difficulties, and the ratio of standard deviations (SD) between the two calibrations. For consistent item performance, the critical values for comparing the robust *Z* statistic is 2.7, the correlations should be at least 0.95, and the ratio of standard deviations should be between 0.90 and 1.10 (Huynh & Meyer, 2010). The outlier in Figure 6.1.1 is detected by a robust *Z* statistic greater than a critical value of 2.7. Table 6.1.1 reports these statistics of correlations and SD ratio for the 2014NeSA-AAR test. Not surprisingly, the two statistics (0.86 for the correlation and 0.80 for SD ratio) for Grade 11 fall below the desired ranges, probably due to the existence of that outlier item seen above on the plot. Appendices N – P presented these statistics for each grade and content area.

Table 6.1.1 NeSA-AAR Pre- and Post-Equating Comparison

	Grade						
	3	4	5	6	7	8	11
Correlation	0.92	0.92	0.84	0.84	0.92	0.93	0.86
SD pre	0.63	0.61	0.51	0.45	0.80	0.80	0.59
SD post	0.80	0.61	0.53	0.54	0.88	0.82	0.74
SD Ratio	0.79	1.00	0.96	0.84	0.92	0.97	0.80

For all content areas, items that departed significantly from the linear trend or below the ideal ranges of robust *Z*, correlation, or SD ratio values were further evaluated by the NDE to determine whether to include those items in the linking set for equating. After evaluating the evidence for the stability between the old (banked) and new (2014) item data and possible item content changes, the NDE decided to drop that item from the linking set used for the 2014 NeSA-AAR Grade 11 test equating.

As a summary of the 2014 NeSA-Alt test equating solutions, NDE decided to adopt a post-equating process for NeSA-AAR Grade 7, NeSA-AAR Grade 11, NeSA-AAM Grade 3, NeSA-AAM Grade 11, and NeSA-AAS Grade 11. For these tests, test difficulty was adjusted by excluding the outlier items and then applied to the raw-to-scale-score conversion. For the other grade and content areas, NDE decided to use a pre-equating solution, keep the whole set of operational items in the linking set and then apply to the raw-to-scale-score conversion. As an additional protective measure, any item that is dropped from either the test form or the equating is excluded from use on future forms.

6.2 SCALING

The purpose of a scaling analysis is to create a score scale. The basic score on any test is the raw score, which is the number of items answered correctly or the total score points earned. However, the raw score alone does not present a wide-ranging picture of test performance because it is not on an equal-interval scale and can be interpreted only in terms of a particular set of items. Since a given raw score may not represent the same skill level on every test form, scale scores were assigned to each raw score point to adjust for slight shifts in item difficulties and permit valid comparison across all test administrations within a particular content area.

Defining the scale score metric is an important, albeit arbitrary, step. Mathematically, scale scores are a linear transformation of the logit scores and thus do not alter the relationships or the displays. Scale scores are the numbers that will be reported to describe the performance of the students, schools, and systems. They will define the ranges of the performance levels, appear on individual student reports and school accountability analyses, and be dissected in newspaper accounts.

Appendix Q contains the detailed raw-score-to-scale-score conversion tables that were used to assign scale scores to students based on the total number correct scores from the NeSA-AAR for 2014, Appendix R for NeSA-AAM for 2014 and Appendix S for NeSA-AAS 2014. Because the relationship between raw and scale scores depends on the difficulties of the specific items on the form, these tables will change for every operational form.

There are two primary considerations when establishing the metric:

- Multiply the logit by a value large enough to make decimal points unnecessary for student scores, and
- Shift the scale enough to avoid negative values for low scale scores.

The scale chosen, for all grades and content areas of the NeSA-Alt assessment, range from 0 to 200. The value of 0 is reserved for students who were not tested or were otherwise invalidated. Thus, any student who attempted the test will receive a scale score equal to 1 even if the student gave no correct responses. No student tested will receive a scale score higher than 200 or lower than 1, even if this requires constraining the scale score calculation. It is possible that a future form will be easy enough that the upper limit of 200 is not invoked even for a perfect paper or could be difficult enough that the lower limit is not invoked.

As part of its deliberations concerning defining the performance levels, the State Board of Education specified that the *Meets the Standards* performance level have a scale score of 85 and that the *Exceeds the Standards* level have a scale score of 135. The logit standards defining the performance levels were adopted by the State Board of Education per the standard setting.

Complete documentation of all standard setting events are presented in separate documents and are placed on the Nebraska State Department of Education website labeled:

2010 NeSA-AAR Standard Setting Technical Report, http://www.education.ne.gov/Assessment/pdfs/2010_NeSA_AAR_Standard_Setting_Report.pdf

2011 NeSA-AAR and NeSA-AAM Standard Setting Technical Report, http://www.education.ne.gov/Assessment/pdfs/2011_NeSA_AAR_and_AAM_Standard_Setting_Report.pdf

and *2012 NeSA-AAS Standard Setting Technical Report*, http://www.education.ne.gov/Assessment/pdfs/NeSA-AAS%20Standard%20Setting%20Results.pdf

Given the scale score and the logit standards defining the performance level, it is sufficient to define the final scale score metric. To ensure proper rounding on all future forms, the calculations used 84.501 and 134.501 as the scale score performance standards. The arithmetic was done using logits rounded to four decimals and the final constants for the slope and intercept of the transformation were rounded to five. Scale scores are rounded to whole numbers.

The transformation to scale scores is:

1. SS = $a + b *$ logit where
2. $b = \frac{134.501 - 84.501}{x_E - x_M}$ where x_E is the logit for *Exceeds Standards* and x_M is the logit for *Meets Standards*.
3. $a = 84.501 - bx_M$ or $a = 134.501 - bx_E$.

Calculations of the slopes and intercepts for all grades of the NeSA-AAR scale score conversion are given in Table 6.2.1, for NeSA-AAM 6.2.2, and for NeSA-AAS 6.2.3. The raw-to-scale conversions are provided in Appendices Q, R, and S.

Table 6.2.1 NeSA-AAR Conversion of Logits to Scale Scores

	Logit Cut Points		Scale Score Ranges by Performance Level			Conversion	
Grade	B/M	M/E	Below	Meets	Exceeds	Slope b	Intercept a
3	0.2501	1.8426	1 to 84	85-134	135 to 200	31.39720	76.64840
4	0.2536	1.8106	1 to 84	85-134	135 to 200	32.11300	76.35520
5	0.2612	1.5392	1 to 84	85-134	135 to 200	39.12360	74.28010
6	0.4202	2.0909	1 to 84	85-134	135 to 200	29.92760	71.92460
7	0.4169	1.7456	1 to 84	85-134	135 to 200	37.63080	68.81120
8	0.6792	2.3138	1 to 84	85-134	135 to 200	30.58680	63.72690
11	0.2362	1.8139	1 to 84	85-134	135 to 200	31.69170	77.01370

Table 6.2.2 NeSA-AAM Conversion of Logits to Scale Scores

	Logit Cut Points		Scale Score Ranges by Performance Level			Conversion	
Grade	B/M	M/E	Below	Meets	Exceeds	Slope b	Intercept a
3	-0.0819	1.6006	1 to 84	85-134	135 to 200	29.71770	86.93460
4	0.4250	1.7728	1 to 84	85-134	135 to 200	37.09750	68.73270
5	-0.0108	1.3462	1 to 84	85-134	135 to 200	36.84600	84.89680
6	0.2970	2.0591	1 to 84	85-134	135 to 200	28.37520	76.07320
7	0.2953	1.7471	1 to 84	85-134	135 to 200	34.44000	74.33050
8	0.4528	1.7661	1 to 84	85-134	135 to 200	38.07200	67.26220
11	0.2976	1.2809	1 to 84	85-134	135 to 200	50.84920	69.36900

Table 6.2.3 NeSA-AAS Conversion of Logits to Scale Scores

	Logit Cut Points		Scale Score Ranges by Performance Level			Conversion	
Grade	B/M	M/E	Below	Meets	Exceeds	Slope b	Intercept a
5	-1.0631	0.3571	1 to 84	85-134	135 to 200	35.20631	121.93783
8	-0.7286	0.5524	1 to 84	85-134	135 to 200	39.03201	112.94872
11	-0.8043	0.6780	1 to 84	85-134	135 to 200	33.73136	111.64013

Complete frequency distributions of the state scale scores for the NeSA-AAR, NeSA-AAM, and NeSA-AAS are provided in Appendices Q, R, and S as part of the raw-to-scale-score conversion tables. In addition, descriptive statistics of the state raw scores, scale scores, and performance levels are computed for subgroups based on gender, ethnicity, special education status, limited English proficiency status, and food program eligibility status in Appendix T. A simple summary of the reading, mathematics, and science distributions can be found in Tables 6.2.4, 6.2.5, and 6.2.6.

Table 6.2.4 2014 NeSA-AAR State Scale Score Summary, All Students

Grade	Count	Scale Score		Quartile		
		Mean	S.D.	First	Second	Third
3	282	108.7	52.4	75	110	145
4	332	106.0	51.3	76	104	141
5	326	113.2	59.1	75	117	160
6	319	112.4	48.3	79	111	143
7	334	110.0	60.2	63	116	151
8	337	108.5	54.2	72	112	147
11	297	114.4	56.8	73	127	153

Table 6.2.5 2014 NeSA-AAM State Scale Score Summary, All Students

Grade	Count	Scale Score		Quartile		
		Mean	S.D.	First	Second	Third
3	276	107.4	49.8	80	112	146
4	333	103.4	58.2	60	100	147
5	331	113.4	58.7	64	116	157
6	317	105.6	43.4	73	104	135
7	332	104.4	53.7	69	101	151
8	344	99.2	55.0	56	104	135
11	304	99.9	63.9	44	100	157

Table 6.2.6 2014 NeSA-AAS State Scale Score Summary, All Students

Grade	Count	Scale Score		Quartile		
		Mean	S.D.	First	Second	Third
5	325	116.7	59.0	78	124	163
8	333	106.5	59.9	61	110	153
11	296	110.8	56.1	71	118.5	151

7. FIELD TEST ITEM DATA SUMMARY

As noted in Chapter Two, in addition to the operational items, field test items were embedded in all content areas and grade level assessments in order to expand the item pool for future form development. Field test items are items being administered for the first time to gather statistical information. These items do not count toward an individual student's score. All field tested items were analyzed statistically following classical item analysis methods including proportion correct, point-biserial correlation, and DIF.

7.1 CLASSICAL ITEM STATISTICS

Indices known as classical item statistics included the item *p*-value and the point-biserial correlations for MC items. For MC items, the *p*-value reflects the proportion of students who answered the item correctly. In general, more capable students are expected to respond correctly to easy items and less capable students are expected to respond incorrectly to difficult items. The primary way of detecting such conditions is through the point-biserial correlation coefficient for dichotomous (MC) items. The point-biserial correlation will be positive if the total test mean score is higher for the students who respond correctly to MC items and negative when the reverse is true.

The traditional statistics are computed for each NeSA-AAR field test item in Appendix F, for NeSA-AAM Appendix G and NeSA-AAS Appendix H. Tables 7.1.1, 7.1.2, and 7.1.3 provide summaries of the distributions of item proportion correct and point-biserial correlations. For future form construction, items with negative point-biserial correlations are never considered for operational use. Items with correlations less than 0.2 or proportion correct less than 0.3 or greater 0.9 are avoided when possible.

Table 7.1.1 Summary of Traditional Item Statistics for NeSA-AAR 2014 Field Test Items

	Item Proportion Correct											
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	Mean	Total
3	0	0	1	2	1	2	5	3	2	0	0.596	16
4	0	0	0	0	3	5	6	1	1	0	0.610	16
5	0	0	0	1	1	4	5	3	2	0	0.642	16
6	0	0	0	0	1	1	5	6	3	0	0.711	16
7	0	0	0	0	0	2	8	5	1	0	0.678	16
8	0	0	0	2	2	4	3	4	1	0	0.606	16
11	0	0	0	3	1	0	5	5	2	0	0.645	16

	Item Point-biserial Correlation							
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	Total
3	0	0	1	2	5	3	5	16
4	0	0	0	0	4	9	3	16
5	0	0	2	1	2	2	9	16
6	0	0	0	2	4	5	5	16
7	0	0	0	1	4	2	9	16
8	0	0	1	1	3	5	6	16
11	0	1	2	0	2	6	5	16

Table 7.1.2 Summary of Traditional Item Statistics for NeSA-AAM 2014 Field Test Items

	Item Proportion Correct											
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	Mean	Total
3	0	0	2	0	5	5	2	2	0	0	0.518	16
4	0	0	0	1	4	7	1	2	1	0	0.557	16
5	0	0	1	4	4	4	2	1	0	0	0.489	16
6	0	0	0	5	1	5	3	0	2	0	0.538	16
7	0	0	1	2	4	3	2	3	1	0	0.553	16
8	0	0	2	0	3	4	3	4	0	0	0.569	16
11	0	0	1	3	4	5	1	1	1	0	0.498	16

	Item Point-biserial Correlation							
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	Total
3	0	0	2	1	2	7	4	16
4	0	0	0	2	5	4	5	16
5	1	3	0	4	2	4	2	16
6	0	0	2	5	0	4	5	16
7	1	0	4	1	1	3	6	16
8	0	0	3	2	1	7	3	16
11	3	0	2	3	4	3	1	16

Table 7.1.3 Summary of Traditional Item Statistics for NeSA-AAS 2014 Field Test Items

	Item Proportion Correct											
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	Mean	Total
5	0	0	0	0	1	3	5	6	1	0	0.672	16
8	0	0	0	1	1	4	4	3	3	0	0.655	16
11	0	0	0	0	0	3	6	5	2	0	0.690	16

	Item Point-biserial Correlation							
Grade	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	Total
5	0	0	0	2	2	4	8	16
8	1	0	0	0	6	2	7	16
11	0	0	0	1	3	3	9	16

8. RELIABILITY

This chapter addresses the reliability of NeSA-Alt test scores. According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), reliability refers to

> the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group (p. 25).

8.1 COEFFICIENT ALPHA

The ability to measure consistently is a necessary prerequisite for making appropriate interpretations (i.e., showing evidence of valid use of results). Conceptually, reliability can be referred to as the consistency of the results between two measures of the same thing. This consistency can be seen in the degree of agreement between two measures on two occasions. Operationally, such comparisons are the essence of the mathematically defined reliability indices.

All measures consist of an accurate, or true, component and an inaccurate, or error, component. Errors occur as a natural part of the measurement process and can never be eliminated entirely. For example, uncontrollable factors such as differences in the physical environment and changes in examinee disposition may increase error and decrease reliability. This is the fundamental premise of traditional reliability analysis and measurement theory. Stated explicitly, this relationship can be seen as the following:

$$\textit{Observed Score} = \textit{True Score} + \textit{Error} \qquad (8.1)$$

To facilitate a mathematical definition of reliability, these components can be rearranged to form the following ratio:

$$\textit{Reliability} = \frac{\textbf{\textit{TrueScoreVariance}}}{\textbf{\textit{ObservedScoreVariance}}} = \frac{\textbf{\textit{TrueScoreVariance}}}{\textbf{\textit{TrueScoreVariance}} + \textbf{\textit{ErrorScore Variance}}} \qquad (8.2)$$

When there is no error, the reliability is true score variance divided by true score variance, which equals 1. However, as more error influences the measure, the error component in the denominator of the ratio increases. As a result, the reliability decreases.

The reliability index used for the 2014 administration of the NeSA-Alt was the Coefficient Alpha α (Cronbach, 1951). Acceptable α values generally range in the mid to high 0.80s to low 0.90s. The total test Coefficient Alpha reliabilities of the whole population are presented in Table 8.1.1 for each grade and content area of the NeSA-Alt. The table contains test length in total number of items (*L*), test reliabilities, and traditional standard errors of measurement (SEM). As can be seen in the table, all reading, mathematics, and science forms for grades 3-11 have Coefficient Alphas in the low 0.90s. Overall, these α values provide evidence of good reliability.

Table 8.1.1 Reliabilities and Standard Errors of Measurement

	Grade	L	Reliability	SEM
Reading	3	282	0.93	1.9
	4	332	0.92	2.0
	5	326	0.93	1.9
	6	319	0.91	2.0
	7	334	0.93	1.8
	8	337	0.94	1.9
	11	297	0.94	1.7
Mathematics	3	276	0.92	1.9
	4	333	0.94	2.1
	5	331	0.94	2.1
	6	317	0.91	2.2
	7	332	0.94	2.1
	8	344	0.93	2.1
	11	304	0.93	2.1
Science	5	325	0.93	1.9
	8	333	0.93	1.9
	11	296	0.95	2.0

Appendix U present α for the content strands. Given that α is a function of test length, the smaller item counts for the content standards result in lower values of α which is to be expected. Reliability estimates for subgroups based on gender, ethnicity, special education status, limited English proficiency status, and food program eligibility status are not computed for the NeSA-Alt tests due to the small sample size of some subgroups.

8.2 STANDARD ERROR OF MEASUREMENT

The traditional SEM uses the information from the test along with an estimate of reliability to make statements about the degree to which error influences individual scores. The SEM is based on the premise that underlying traits, such as academic achievement, cannot be measured exactly without a perfectly precise measuring instrument. The standard error expresses unreliability in terms of the raw-score metric. The SEM formula is provided below:

$$SEM = SD\sqrt{1 - reliability} \qquad (8.3)$$

This formula indicates that the value of the SEM depends on both the reliability coefficient and the standard deviation of test scores. If the reliability were equal to 0.00 (the lowest possible value), the SEM would be equal to the standard deviation of the test scores. If test reliability were equal to 1.00 (the highest possible value), the SEM would be 0.0. In other words, a perfectly reliable test has no measurement error (Harvill, 1991). SEMs were calculated for each NeSA-Alt grade and content area using raw scores and displayed in Table 8.1.1.

8.3 CONDITIONAL STANDARD ERROR OF MEASUREMENT (CSEM)

The preceding discussion reviews the traditional approach to judging a test's consistency. This approach is useful for making overall comparisons between alternate forms. However, it is not very useful for judging the precision with which a specific student's score is known. The Rasch measurement models provide "conditional standard errors" that pertain to each unique ability estimate. Therefore, the CSEM may be especially useful in characterizing measurement precision in the neighborhood of a score level used for decision-making—such as cut scores for identifying students who meet a performance standard.

The complete set of conditional standard errors for every obtainable score can be found in Appendices Q, R and S as part of the raw-to-scale-score conversions for each grade and content area. Values were derived using the calibration data file described in Chapter Six and are on the scaled score metric. The magnitudes of CSEMs across the score scale seemed reasonable for most NeSA-Alt tests that the values are lower in the middle of the score range and increase at both extremes (i.e., at smaller and larger scale scores). This is because ability estimates from scores near the center of the test scoring range are known much more precisely than abilities associated with extremely high or extremely low scores. Table 8.3.1 reports the minimum CSEM of the scale score associated with the zero total test score (Min CSEM), the maximum CSEM of the scale score associated with the perfect total test score (Max CSEM), CSEM at the cuts of Below and Meets performance levels (CSEM B/M), and CSEM at the cuts of Meets and Exceeds performance levels (CSEM M/E) for each grade and content area. CSEM values at the cut score were generally associated with smaller CSEM values, indicating that more precise measurement occurs at these cuts.

Table 8.3.1 CSEM of the Scale Scores for 2014 NeSA-Alt Tests

	Grade	Min CSEM	Max CSEM	CSEM B/M	CSEM M/E
Reading	3	13	58	13	18
	4	13	59	13	17
	5	16	72	16	20
	6	12	55	12	17
	7	16	69	16	21
	8	13	56	13	17
	11	13	58	14	20
Mathematics	3	13	55	13	17
	4	14	68	14	19
	5	14	68	14	18
	6	11	52	11	16
	7	13	63	14	19
	8	15	70	15	19
	11	20	94	20	24
Science	5	15	65	15	17
	8	16	72	16	22
	11	13	62	13	19

8.4 DECISION CONSISTENCY AND ACCURACY

When criterion-referenced tests are used to place the examinees into two or more performance classifications, it is useful to have some indication of how accurate or consistent such classifications are. Decision consistency refers to the degree to which the achievement level for each student can be replicated upon retesting using an equivalent form (Huynh, 1976). Decision accuracy describes the extent to which achievement-level classification decisions based on the administered test form would agree with the decisions that would be made on the basis of a perfectly reliable test. In a standards-based testing program there should be great interest in knowing how consistently and accurately students are classified into performance categories.

Since it is not feasible to repeat NeSA-Alt testing in order to estimate the proportion of students who would be reclassified in the same achievement levels, a statistical model needs to be imposed on the data to project the consistency or accuracy of classifications solely using data from the available administration (Hambleton & Novick, 1973). Although a number of procedures are available, two well-known methods were developed by Hanson and Brennan (1990) and Livingston and Lewis (1995) utilizing specific True Score Models. These approaches are fairly complex, and the cited sources contain details regarding the statistical models used to calculate decision consistency from the single NeSA-Alt administration.

Several factors might affect decision consistency. One important factor is the reliability of the scores. All other things being equal, more reliable test scores tend to result in more similar reclassifications. Another factor is the location of the cutscore in the score distribution. More consistent classifications are observed when the cutscores are located away from the mass of the score distribution. The number of performance levels is also a consideration. Consistency indices for four performance levels should be lower than those based on three categories because classification using four levels would allow more opportunity to change achievement levels. Finally, some research has found that results from the Hanson and Brennan (1990) method on a dichotomized version of a complex assessment yield similar results to the Livingston and Lewis method (1995) and the method by Stearns and Smith (2007).

The results for the overall consistency across all three achievement levels are presented in Tables 8.4.1 – 8.4.3. The tabled values, derived using the program *BB-Class* (Brennan, 2004), show that consistency values across the two methods are generally very similar. Across all content areas, the overall decision consistency ranged from the mid 0.80s to the low 0.90s while the decision accuracy ranged from the high 0.80s to the mid 0.90s. If a parallel test were administered, at least 85% or more of students would be classified in the same way. Dichotomous decisions using the Meets cuts (Below/Meets) generally have the highest consistency values and exceeded 0.90 in all cases. The pattern of decision accuracy across different cuts is similar to that of decision consistency.

Table 8.4.1 NeSA-AAR Decision Consistency Results

Content Area	Grade	Livingston & Lewis				Hanson & Brennan			
		Decision Accuracy		Decision Consistency		Decision Accuracy		Decision Consistency	
		Meets	Exceeds	Meets	Exceeds	Meets	Exceeds	Meets	Exceeds
Reading	3	0.94	0.92	0.92	0.89	0.94	0.91	0.92	0.89
	4	0.93	0.92	0.90	0.89	0.93	0.92	0.90	0.89
	5	0.90	0.88	0.86	0.83	0.91	0.90	0.88	0.86
	6	0.93	0.92	0.90	0.88	0.93	0.92	0.90	0.89
	7	0.94	0.91	0.91	0.88	0.94	0.91	0.92	0.89
	8	0.95	0.93	0.92	0.90	0.94	0.92	0.92	0.90
	11	0.95	0.92	0.93	0.88	0.95	0.92	0.93	0.89

Table 8.4.2 NeSA-AAM Decision Consistency Results

Content Area	Grade	Livingston & Lewis				Hanson & Brennan			
		Decision Accuracy		Decision Consistency		Decision Accuracy		Decision Consistency	
		Meets	Exceeds	Meets	Exceeds	Meets	Exceeds	Meets	Exceeds
Math	3	0.94	0.89	0.92	0.85	0.94	0.89	0.92	0.86
	4	0.94	0.93	0.91	0.91	0.94	0.93	0.91	0.91
	5	0.94	0.93	0.92	0.90	0.94	0.93	0.92	0.91
	6	0.92	0.92	0.89	0.89	0.92	0.92	0.89	0.89
	7	0.94	0.93	0.92	0.91	0.94	0.93	0.92	0.91
	8	0.93	0.92	0.90	0.89	0.93	0.92	0.90	0.89
	11	0.92	0.92	0.89	0.89	0.92	0.92	0.89	0.89

Table 8.4.3 NeSA-AAS Decision Consistency Results

Content Area	Grade	Livingston & Lewis				Hanson & Brennan			
		Decision Accuracy		Decision Consistency		Decision Accuracy		Decision Consistency	
		Meets	Exceeds	Meets	Exceeds	Meets	Exceeds	Meets	Exceeds
Science	5	0.94	0.93	0.91	0.90	0.94	0.92	0.92	0.90
	8	0.93	0.92	0.91	0.89	0.93	0.92	0.91	0.89
	11	0.96	0.92	0.94	0.89	0.95	0.92	0.94	0.90

9. VALIDITY

As defined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), validity refers to “the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests” (p. 9). The validity process involves the collection of a variety of evidence to support the proposed test score interpretations and uses. This entire technical report describes the technical aspects of the NeSA-Alt tests in support of their score interpretations and uses. Each of the previous chapters contributes important evidence components that pertain to score validation: test development, test scoring, item analysis, Rasch calibration, scaling, and reliability. This chapter summarizes and synthesizes the evidence based on the framework presented in *The Standards*.

9.1 EVIDENCE BASED ON TEST CONTENT

Content validity addresses whether the test adequately samples the relevant material it purports to cover. The NeSA-Alt for grades 3 through 11 is a criterion-referenced assessment. The criteria referenced are the Nebraska reading and mathematics content standards. Each assessment was based on and was directly aligned to the Nebraska statewide alternate content standards to ensure good content validity.

For criterion-referenced, standards-based assessment, the strong content validity evidence is derived directly from the test construction process and the item scaling. The item development and test construction process, described above, ensures that every item aligns directly to one of the content standards. This alignment is foremost in the minds of the item writers and editors. As a routine part of item selection prior to an item appearing on a test form, the review committees check the alignment of the items with the standards and make any adjustments necessary. The result is consensus among the content specialists and teachers that the assessment does in fact assess what was intended.

The empirical item scaling, which indicates where each item falls on the logit ability-difficulty continuum, should be consistent with what theory suggests about the items. Items that require more knowledge, more advanced skills, and more complex behaviors should be empirically more difficult than those requiring less. Evidence of this agreement is contained in the item summary tables in Appendices K, L, and M.

9.2 EVIDENCE BASED ON INTERNAL STRUCTURE

As described in the *Standards* (1999), internal-structure evidence refers to the degree to which the relationships between test items and test components conform to the construct on which the proposed test interpretations are based.

Item-Test Correlations: Item-test correlations are reviewed in Chapter Four. All values are positive and of acceptable magnitude.

Item Response Theory Dimensionality: Results from principle components analyses are presented in Chapter Five. The NeSA-Alt reading, mathematics, and science tests were essentially unidimensional, providing evidence supporting interpretations based on the total scores for the respective NeSA-Alt tests.

Strand Correlations: Correlations and disattenuated correlations between strand scores within each content area are presented below. This data can also provide information on score dimensionality that is part of internal-structure evidence. As noted in Chapter Two and also in Table 9.2.1, the NeSA-AAR tests have two strands (denoted by R.1 and R.2), the NeSA-AAM tests have four strands (denoted by M.1, M.2, M.3, and M.4), and the NeSA-AAS have four strands (denoted by S.1, S.2, S.3, and S.4) for each grade and content area.

For each grade, Pearson's correlation coefficients between these strands are reported in Tables 9.2.2.a through 9.2.2.g. The intercorrelations between the strands within the content areas are positive and generally range from moderate to high in value.

Table 9.2.1 NeSA-Alt Content Strands

Content	Code	Strand
Reading	R.1	Vocabulary
	R.2	Comprehension
Mathematics	M.1	Number Sense
	M.2	Geometric/Measurement
	M.3	Algebraic
	M.4	Data Analysis/Probability
Science	S.1	Inquiry, the Nature of Science, and Technology
	S.2	Physical Science
	S.3	Life Science
	S.4	Earth and Space Science

Table 9.2.2.a Correlations between Reading and Mathematics Strands for Grade 3

Grade 3	R.1	R.2	M.1	M.2	M.3	M.4
R.1						
R.2	0.85					
M.1	0.84	0.84				
M.2	0.83	0.80	0.81			
M.3	0.73	0.76	0.71	0.68		
M.4	0.73	0.77	0.72	0.72	0.65	

Table 9.2.2.b Correlations between Reading and Mathematics Strands for Grade 4

Grade 4	R.1	R.2	M.1	M.2	M.3	M.4
R.1						
R.2	0.84					
M.1	0.82	0.84				
M.2	0.81	0.84	0.88			
M.3	0.75	0.80	0.80	0.80		
M.4	0.65	0.66	0.69	0.69	0.65	

Table 9.2.2.c Correlations between Reading, Mathematics, and Science Strands for Grade 5

Grade 5	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1										
R.2	0.85									
M.1	0.83	0.87								
M.2	0.81	0.81	0.87							
M.3	0.68	0.73	0.77	0.70						
M.4	0.66	0.69	0.70	0.64	0.53					
S.1	0.71	0.77	0.79	0.72	0.67	0.66				
S.2	0.77	0.82	0.80	0.76	0.70	0.55	0.72			
S.3	0.81	0.84	0.83	0.79	0.67	0.59	0.72	0.80		
S.4	0.81	0.83	0.80	0.79	0.65	0.65	0.74	0.78	0.80	

Table 9.2.2.d Correlations between Reading and Mathematics Strands for Grade 6

Grade 6	R.1	R.2	M.1	M.2	M.3	M.4
R.1						
R.2	0.82					
M.1	0.72	0.78				
M.2	0.73	0.80	0.73			
M.3	0.70	0.73	0.76	0.71		
M.4	0.68	0.72	0.71	0.72	0.66	

Table 9.2.2.e Correlations between Reading and Mathematics Strands for Grade 7

Grade 7	R.1	R.2	M.1	M.2	M.3	M.4
R.1						
R.2	0.83					
M.1	0.77	0.83				
M.2	0.74	0.78	0.78			
M.3	0.67	0.71	0.71	0.71		
M.4	0.81	0.88	0.83	0.75	0.73	

Table 9.2.2.f Correlations between Reading, Mathematics, and Science Strands for Grade 8

Grade 8	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1										
R.2	0.85									
M.1	0.73	0.75								
M.2	0.74	0.80	0.78							
M.3	0.78	0.81	0.78	0.76						
M.4	0.81	0.82	0.78	0.77	0.80					
S.1	0.74	0.79	0.72	0.78	0.75	0.76				
S.2	0.78	0.82	0.73	0.75	0.79	0.79	0.78			
S.3	0.81	0.86	0.73	0.78	0.77	0.82	0.77	0.81		
S.4	0.71	0.73	0.69	0.70	0.69	0.73	0.66	0.70	0.70	

Table 9.2.2.g Correlations between Reading, Mathematics, and Science Strands for Grade 11

Grade 11	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1										
R.2	0.85									
M.1	0.65	0.75								
M.2	0.77	0.83	0.76							
M.3	0.81	0.84	0.71	0.81						
M.4	0.64	0.71	0.64	0.69	0.74					
S.1	0.63	0.68	0.62	0.63	0.65	0.54				
S.2	0.77	0.82	0.68	0.75	0.75	0.67	0.70			
S.3	0.81	0.86	0.69	0.78	0.82	0.66	0.69	0.85		
S.4	0.79	0.82	0.63	0.75	0.79	0.63	0.66	0.82	0.87	

The correlations in Tables 9.2.2.a through 9.2.2.g are based on the observed strand scores. These observed-score correlations are weakened by existing measurement error contained within each strand. As a result, disattenuating the observed correlations can provide an estimate of the relationships between strands if there is no measurement error. The disattenuated correlation coefficients can be computed from the observed correlations (reported in Tables 9.2.2.a – 9.2.2.g) and the reliabilities for each strand (Spearman, 1904, 1910). Disattenuated correlations very near 1.00 might suggest that the same or very similar constructs are being measured. Values somewhat less than 1.00 might suggest that different strands are measuring slightly different aspects of the same construct. Values markedly less than 1.00 might suggest the strands reflect different constructs.

Tables 9.2.3.a through 9.2.3.g show the corresponding disattenuated correlations for the 2014 NeSA-Alt tests for each grade. Given that none of these strands has perfect reliabilities (see Chapter Eight), the disattenuated strand correlations are higher than their observed score counterparts. Some within-content-area correlations are very high (e.g., above 0.95), suggesting that the within-content-area strands might be measuring essentially the same construct. This, in turn, suggests that some strand scores might not provide unique information about the strengths or weaknesses of students.

On a fairly consistent basis, the correlations between the strands within each content area were higher than the correlations between strands across different content areas. In general, within-content-area strand correlations were mostly close to 1.00, while across-content-area strand correlations generally ranged from 0.71 to 0.99. Such a pattern is expected since the two content area tests were designed to measure different constructs.

Table 9.2.3.a Disattenuated Strand Correlations for Reading and Mathematics: Grade 3

Grade 3	R.1	R.2	M.1	M.2	M.3	M.4
R.1						
R.2	0.99					
M.1	1.00	0.98				
M.2	1.00	0.93	1.00			
M.3	0.99	1.00	0.98	0.94		
M.4	0.99	1.00	0.99	0.98	1.00	

Table 9.2.3.b Disattenuated Strand Correlations for Reading and Mathematics: Grade 4

Grade 4	R.1	R.2	M.1	M.2	M.3	M.4
R.1						
R.2	1.00					
M.1	1.00	0.96				
M.2	1.00	0.99	1.00			
M.3	0.96	0.95	0.97	0.97		
M.4	0.99	0.95	1.00	1.00	0.98	

Table 9.2.3.c Disattenuated Strand Correlations for Reading, Mathematics and Science: Grade 5

Grade 5	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1										
R.2	1.00									
M.1	0.98	0.97								
M.2	0.98	0.94	1.00							
M.3	0.93	0.95	1.00	0.95						
M.4	0.85	0.86	0.87	0.83	0.77					
S.1	0.94	0.98	1.00	0.94	0.99	0.92				
S.2	0.93	0.95	0.93	0.91	0.94	0.71	0.94			
S.3	1.00	0.99	0.98	0.97	0.93	0.78	0.97	0.98		
S.4	1.00	0.98	0.94	0.96	0.90	0.85	0.98	0.95	1.00	

Table 9.2.3.d Disattenuated Strand Correlations for Reading and Mathematics: Grade 6

Grade 6	R.1	R.2	M.1	M.2	M.3	M.4
R.1						
R.2	1.00					
M.1	0.96	0.94				
M.2	0.96	0.96	0.93			
M.3	1.00	0.94	1.00	0.99		
M.4	0.99	0.96	1.00	1.00	1.00	

Table 9.2.3.e Disattenuated Strand Correlations for Reading and Mathematics: Grade 7

Grade 7	R.1	R.2	M.1	M.2	M.3	M.4
R.1						
R.2	1.00					
M.1	0.99	0.97				
M.2	0.94	0.91	0.97			
M.3	0.98	0.94	1.00	1.00		
M.4	1.00	0.99	1.00	0.91	1.00	

Table 9.2.3.f Disattenuated Strand Correlations for Reading, Mathematics and Science: Grade 8

Grade 8	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1										
R.2	0.96									
M.1	0.89	0.91								
M.2	0.88	0.94	1.00							
M.3	0.99	1.01	1.05	1.02						
M.4	0.95	0.95	0.98	0.95	1.04					
S.1	0.94	0.99	0.98	1.03	1.06	0.99				
S.2	0.96	0.99	0.96	0.97	1.08	1.00	1.06			
S.3	0.93	0.97	0.89	0.93	0.99	0.97	0.99	1.00		
S.4	0.93	0.95	0.97	0.96	1.01	0.98	0.97	0.98	0.93	

Table 9.2.3.g Disattenuated Strand Correlations for Reading, Mathematics and Science: Grade 11

Grade 11	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1										
R.2	0.98									
M.1	0.84	0.90								
M.2	0.94	0.96	0.97							
M.3	0.98	0.95	0.90	0.98						
M.4	0.93	0.98	0.98	1.00	1.00					
S.1	0.89	0.90	0.91	0.88	0.90	0.91				
S.2	0.94	0.94	0.86	0.90	0.90	0.98	0.97			
S.3	0.95	0.96	0.85	0.92	0.95	0.94	0.93	0.99		
S.4	0.96	0.93	0.81	0.91	0.95	0.92	0.92	0.98	1.00	

9.3 EVIDENCE RELATED TO THE USE OF THE RASCH MODEL

Since the Rasch model is the basis of all calibration, scaling, and linking analyses associated with the NeSA-Alt, the validity of the inferences from these results depends on the degree to which the assumptions of the model are met as well as the fit between the model and test data. As discussed at length in Chapter Five, the underlying assumptions of Rasch models were essentially met for all the NeSA-Alt data, indicating the appropriateness of using the Rasch models to analyze the NeSA-Alt data.

In addition, the Rasch model was also used to link different operational NeSA-Alt tests across years. The accuracy of the linking also affects the accuracy of student scores and the validity of score uses. DRC Psychometric Services staffers conducted verifications to check the accuracy of the procedures, including item calibration, conversions from the raw score to the Rasch ability estimate, and conversions from the Rasch ability estimates to the scale scores.

10. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69-81.

Andrich, D. (1988). *Rasch models for measurement*. Newberry Park, CA: Sage Publications.

Brennan, R. (2004). BB-Class (Version 1.0). [Computer Software] Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement & Assessment. CASMA: education.uiowa.edu/casma.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Fischer, G. & Molenaar, I. (1995). *Rasch models – Foundations, recent developments, and applications.* New York: Springer.

Hambleton, R. & Novick, M. (1973). "Toward an Integration of Theory and Method for Criterion-Referenced Tests." *Journal of Educational Measurement*, *10*, 159–170.

Hanson, B. A. & Brennan, R. L. (1990). An Investigation of Classification Consistency Indexes Estimated Under Alternative Strong True Score Theory Models. *Journal of Educational Measurement, 27(4),* 345–359.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practices*, *10*(2), 33–41.

Huynh, H. (1976). "On the Reliability of Decisions in Domain-Referenced Testing." *Journal of Educational Measurement*, 13, 253–264.

Huynh, H. (2000, June). *Guidelines for Rasch Linking for PACT*. Memorandum to Paul Sandifer on June 18, 2000. Columbia, SC: Available from Author.

Huynh, H, & Rawls, A. (2009). A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In Everett V. Smith Jr. & Greg E. Stone (Eds.) *Applications of Rasch Measurement in Criterion-Referenced Testing: Practice Analysis to Score Reporting*. Maple Grove, MN: JAM Press.

Huynh, H. & Meyer, P. (2010). "Use of Robust z in Detecting Unstable Items in Item Response Theory Models." *Practical Research, Assessment, and Evaluation*, *vol 15*, no. 2.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-Based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Linacre, J. M. (2009). *A user's guide to WINSTEPS MININSTEP Rasch-model computer programs*. Chicago, IL: Winsteps.

Linacre, J. M. (2014). *Winsteps® Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com

Livingston, S. & Lewis, C. (1995). "Estimating the Consistency and Accuracy of Classifications Based on Test Scores." *Journal of Educational Measurement 32*, 179–197.

Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement*, *10*, 217–229.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioral Research*, *14*, 21–38.

Mead, R. J. (1976). *Assessing the fit of data to the Rasch model through the analysis of residuals*. Unpublished doctoral dissertation, Chicago: University of Chicago.

Mead, R. J. (2008). *A Rasch primer: The measurement theory of Georg Rasch*. (Psychometrics services research memorandum 2008–001). Maple Grove, MN: Data Recognition Corporation.

Mogilner, A. (1992). *Children's Writer's World Book*. Cincinnati, OH: Writer's Digest Books.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy 14*, 58-94.

Spearman C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Spearman C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *1*, 199–218.

Smith, E. V. & Smith, R. M. (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.

Stearns, M. & Smith R. M. (2007). *Estimation of classification consistency indices for complex assessments: Model based approaches*. Paper presented at the 2007 Annual Convention of the American Educational Research Association. Chicago, IL.

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Brisner, E. P. (1989). EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies. Orlando, FL: Steck-Vaughn Company.

Thompson, S., Johnston, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments. National Center on Educational Outcomes Synthesis Report 44. Minneapolis, MN: University of Minnesota.

Webb, N. L. (2002). Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessment for Four States. Washington, D.C.: Council of Chief State School Officers.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

Wright, B. & Mok, M. (2004). An overview of the family of Rasch measurement models. In E. Smith & R. Smith (Eds.) *Introduction to Rasch measurement. Theory, models and applications*. (pp. 25-47) Maple Grove, MN: JAM Press.

Wright, B. & Panchapakesan, N. (1969). A procedure of sample-free item analysis. *Educational and Psychological Measurement 29,* pp. 23-48.

Wright, B., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.